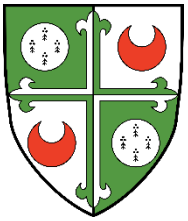
The use of whole exome sequencing data to identify candidate genes involved in cancer and benign tumour predisposition

Eleanor Rose Fewings

Girton College Cambridge



Submission date: July 2018

This dissertation is submitted for the degree of Doctor of Philosophy

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgements

I would like to give special thanks to Marc Tischkowitz who has been an outstanding supervisor. He has provided me continuous opportunities to grow and develop as a researcher and as a person. Most importantly, he has had confidence in my skills when I have been lacking and pushed me to aim high for my future, for which I am extremely grateful.

I also wish to thank my second supervisor Eamonn Maher for his continued advice throughout this project, and Lynda Smith for her unwavering support.

I wish to thank my family and friends who have kept me sane for the past three years. I would particularly like to thank my parents and sister for all their love and for constantly telling me not to work so hard. To Chris Duncan and Charlotte Taylor, thank you for the many hours you listened to me. And finally, I would like to thank Peter Flint for teaching me how to dream big.

1 Table of Contents

Terminology and Glossary

This study uses the term ‘variant’ to describe any sequence change that differs from the reference genome. This includes heterozygous (unless specifically stated that the variant is homozygous) single nucleotide variants, insertions and deletions. Unless otherwise specified, combined counts of variants across genes or terms represent the sum of the allele counts of identified variants, where a heterozygous alternative variant is equal to one, and a homozygous alternative variant is equal to two.

The term ‘protein-affecting variants’ will be used to describe loss of function, inframe indels and predicted deleterious and probably damaging missenses (as flagged by SIFT and PolyPhen respectively; variants with a missing prediction by either SIFT or PolyPhen but otherwise meet the above criteria were also included).

Term	Definition
AC	Allele count
ACC	Adrenocortical carcinoma
AD	Allelic depth
ADP	Adenosine diphosphate
AF	Allele frequency
ALT	Alternative allele to reference
ATP	Adenosine triphosphate
BAM	Binary alignment mapping
BC	Breast cancer
BCC	Basal cell carcinoma
BCL	Base call file format
BQSR	Base quality score recalibration
BWA	Burrows-Wheeler Aligner
cAMP	Cyclic adenosine monophosphate
CAST	Cohort allelic sums test
CBC	Contralateral breast cancer
CDS	Coding sequence
cGMP	cyclic guanosine monophosphate
CMC	Combined multivariate and collapsing
CNV	Copy number variant

COSMIC	Catalogue of somatic mutations in cancer
CSD3	Cambridge service for data driven discovery
CT	Computed tomography
DGC	Diffuse gastric cancer
dNTP	Deoxynucleotide triphosphate
DP	Depth
dTE	Desmoplastic trichoepithelioma
ER	Estrogen receptor
EVE	Exon variability estimate
ExAC	Exome aggregation consortium
FDR	False discovery rate
FFPE	Formalin-fixed paraffin embedded
FPKM	Fragments Per Kilobase of transcript per Million mapped read
GAF	Gene ontology annotation file
GATK	Genome analysis toolkit
gnomAD	Genome aggregation database
GO	Gene ontology
GQ	Genotype quality
GT	Genotype
GWA	Genome wide association
HDGC	Hereditary diffuse gastric cancer
HMM	Hidden Markov model
HNPCC	Hereditary nonpolyposis colorectal cancer
HPA	Human protein atlas
HPC	High performance computing
HR	Hormone receptor
IHC	Immunohistochemistry
IHCAP	Investigating Hereditary Cancer Predisposition
KEGG	Kyoto encyclopedia of genes and genomes
LoH	Loss of heterozygosity
MAC	Microcystic adnexal carcinoma

MALTA	MYH9 associated elastin aggregation
MINAS	multi-locus inherited neoplasia alleles syndrome
MRI	Magnetic resonance imaging
MRPD	MYH9-related platelet disorder
MSI	Microsatellite instability
MT	Mitochondrial
NBS	Nijmegen breakage syndrome
NFE	Non-Finnish European
NGS	Next generation sequencing
NMM	Non-muscle myosin
NMMIIA	Non-muscle myosin II A
NPV	No pathogenic variant
PAV	Protein-affecting variant
PCA	Principle component analysis
PCR	Polymerase chain reaction
PHTS	PTEN Hamartoma Tumour Syndrome
PKA	Protein kinase A
PL	Phred-scaled likelihood
REF	Reference allele
RPKM	Reads Per Kilobase of transcript per Million mapped reads
SCC	Squamous cell carcinoma
SEER	Surveillance, Epidemiology and End Results
SIFT	Sorting intolerant from tolerant
SKAT	Sequence kernel association test
SLURM	Simple Linux Utility for Resource Management
SNP	Single nucleotide polymorphism
SRCC	Signet ring cell carcinoma
TCGA	The cancer genome atlas
TS	Truth sensitivity
UBC	Unilateral breast cancer
UV	Ultraviolet

VAF	Variant allele frequency
VCF	Variant call file
VEP	Variant effect predictor
VQSR	Variant quality score recalibration
WECARE	Women's Environmental Cancer and Radiation Epidemiology
WES	Whole exome sequencing
WGS	Whole genome sequencing
XHMM	Exome-Hidden Markov Mode

Abstract

The development of whole exome sequencing has transformed the study of disease predisposition. The sequencing of both large disease sets and smaller rare disease families enables the identification of new predisposition variants and potentially provide clinical insight into disease management. There is no standard protocol for analysing exome sequencing data. Outside of extremely large sequencing studies including thousands of individuals, statistical approaches are often underpowered to detect rare disease associated variants. Aggregation of variants into functionally related regions, including genes, gene clusters, and pathways, allows for the detection of biological processes that, when interrupted, may impact disease risk. *In silico* functional studies can also be utilised to further understand how variants disrupt biological processes and identify genotype-phenotype relationships.

This study describes the exploration of sequencing datasets from cancers and benign tumour diseases including: i) hereditary diffuse gastric cancer, ii) sweat duct proliferation tumours, iii) adrenocortical carcinoma, and iv) breast cancer. Each set underwent germline whole exome sequencing followed by additional tumour or targeted sequencing to identify associated predisposition genes. Variants within a cluster of risk genes that are involved in double strand break repair were identified as associated with hereditary diffuse gastric cancer risk via gene ontology enrichment analysis. This cluster included *PALB2* within which, using externally collated data, loss of function variants were identified as significantly associated with hereditary diffuse gastric cancer risk. Germline protein-affecting variants in the myosin gene *MYH9* were identified in all individuals with a rare sweat duct proliferative syndrome, suggesting a role for *MYH9* in skin development, regulation and tumorigenesis. These *MYH9* variants were analysed *in silico* to identify a genotype-phenotype relationship between the clinical presentation and variants in the ATP binding pocket of the protein. Sequencing data from adrenocortical carcinoma cases was used to elucidate the role of Lynch syndrome genes in disease pathogenesis. Within the breast cancer set, candidate genes were selected to undergo targeted sequencing in a larger set of cases to further explore their role in breast cancer risk.

Risk associated genes identified within this study may ultimately aid in diagnosis and management of disease. This thesis has also generated multiple novel tools and sequencing analysis techniques that may be of use for further studies by aiding in the prioritisation of candidate variants. The described techniques will provide support to researchers working on rare, statistically underpowered datasets and to provide standard analysis pipelines for a range of dataset sizes and types, including familial data and unrelated individuals.

1 Introduction

1.1 Rare cancer predisposition

Cancer is primarily a genetic disease, arising as a result of an accumulation of genetic and epigenetic abnormalities allowing uninterrupted cell proliferation. Changes that develop within the cells of the affected tissue (somatic variants) are the main drivers behind aberrant cellular proliferation and tumour development. The likelihood of these variants arising can be increased by hereditary defects (germline variants) in key systems such as DNA repair (Stratton, Campbell and Futreal, 2009). Being able to understand the hereditary pathogenic variants that predispose to this oncogenic process can provide means to identify high risk patients, and therefore apply preventative or surveillance strategies. In some cases, it can also provide insight into effective treatment options. For example, PARP inhibitors such as olaparib have been effectively shown to treat women with pathogenic *BRCA1* variants who carry defects in the homologous recombination pathway (Tutt *et al.*, 2010).

Pedigrees describing the pattern of disease in a family can be used to discern how likely that disease is to be caused by hereditary defects, and how those defects have been inherited. The vast majority of familial cancers are inherited in an autosomal dominant manner, meaning that the variant resides on one of the 22 pairs of non-sex chromosomes and that only one copy of the affected gene is required to carry the causal variant (which is therefore heterozygous) for the phenotype to be present. This is in contrast to an autosomal recessive model where the inheritance of two copies of the causal variant (homozygous) is required to produce a phenotype. However these mendelian inheritance patterns overlook issues of penetrance and effect size. Penetrance describes the extent to which a variant causes a particular phenotype, where a penetrant variant produces the associated phenotype in all carriers whereas a low-penetrance variant may only cause the associated phenotype in a small proportion of carriers.

Within breast cancer genetics, many predisposition genes have been identified with varying levels of penetrance. The most well described breast cancer risk genes are *BRCA1* and *BRCA2*. Although these genes are highly penetrant, the occurrence of pathogenic variants is relatively rare and so they are only predicted to account for around 30% of hereditary breast cancer cases (Lalloo and Evans, 2012; Siegel, Naishadham and Jemal, 2013). The Li-Fraumeni syndrome gene *TP53* is also highly penetrant, with pathogenic mutations conferring an increase in risk of 18 to 60 fold, although again these variants are extremely rare and are thought to account for less than 0.1% of all breast cancer cases (Lalloo and Evans, 2012). In contrast, pathogenic variants in moderate penetrance breast cancer susceptibility genes, including *ATM*, *CHEK2*, and *PALB2*, carry an increase in relative risk of around two to four fold (Lalloo and Evans, 2012). Low penetrance variants have been described by genome wide association studies for breast cancer predisposition (Easton *et al.*, 2007; Turnbull, Ahmed and Morrison, 2010; Michailidou *et al.*, 2017). Although the low relative risks associated with these variants are more complicated to

interpret clinically, they may be used to generate polygenic risk scores such as those generated to assess prostate cancer risk (Szulkin *et al.,* 2015; Michailidou *et al.*, 2017).

Cancer risk studies in twins suggests that the proportion of disease risk that is due to heritable factors is greatly dependent on the type of cancer, with inherited genetic factors contributing to 57% of prostate cancers but only 15% of colon cancers (Mucci *et al.*, 2016). Inherited cancers and causal genes are often discovered through segregation and linkage analysis of large pedigrees. One of the first cases of hereditary cancer was identified as far back as 1895 by Dr Aldred Warthin who described one family of 50 individuals of which 18 were affected with cancers of the stomach, uterine, and liver (Warthin, 1913). Since then, the development of genetic analysis has made it possible to identify key predisposition genes, such as the identification of the *BRCA1* gene in breast and ovarian cancer families (Miki *et al.*, 1994) and the discovery of pathogenic variants in E-cadherin gene *CDH1* in a large set of kindred from New Zealand (Aotearoa) affected with diffuse gastric cancer (Guilford *et al.*, 1998).

The identification of highly penetrant variants in cancer genes has greatly impacted disease management strategies. The latest cumulative risk estimates for breast cancer development by 80 years of age for *BRCA1* and *BRCA2* pathogenic variant carriers are 72% and 69% for each gene respectively (Kuchenbaecker *et al.*, 2017). For carriers, management options range from regular magnetic resonance imaging (MRI) surveillance to bilateral risk-reducing mastectomy. For individuals with high lifetime risk estimates, bilateral risk-reducing mastectomy has been shown to substantially improve survival in comparison to surveillance alone (Heemskerk-Gerritsen *et al.*, 2013). As pathogenic variants in *BRCA1* and *BRCA2* also confer a risk to ovarian cancer, a bilateral risk-reducing oophorectomy is also a preventative option for carriers. When performed alongside a risk-reducing mastectomy, a bilateral oophorectomy has been shown to confer no additional risk of postoperative morbidity and so could be a viable option for high risk individuals (Elmi *et al.*, 2018).

Families carrying pathogenic *CDH1* variants have a cumulative risk of diffuse gastric cancer (DGC) development of 70% for men and 56% for women (van der Post *et al.*, 2015). Current guidelines recommend that pathogenic *CDH1* variant carriers undergo risk reducing gastrectomy at the earliest possible point (van der Post *et al.*, 2015). However such an invasive procedure carries a relatively high morbidity rate, with perioperative mortality ranging from 3-6% and risk of fatal complications between 8-15% (Lewis *et al.*, 2001; Norton *et al.*, 2007). Female carriers of pathogenic *CDH1* variants are also at risk of developing lobular breast cancer, with a cumulative risk by the age of 80 years of 39% (Pharoah, Guilford and Caldas, 2001). These individuals are recommended to undergo regular breast cancer surveillance (Pharoah, Guilford and Caldas, 2001). The well-studied risks associated with *CDH1* variants means that carriers can make informed decisions about risk reducing strategies. However, it is predicted that around 60% of hereditary diffuse gastric cancer (HDGC) families are negative for pathogenic *CDH1* variants and so their own personal risk estimates are more difficult to predict.

As well as the problem surrounding cancer families with no known pathogenic variants in the primary disease-causing gene, it is now becoming apparent that pathogenic variants in cancer genes originally considered to be associated with one cancer phenotype have an association with multiple phenotypes. Lynch syndrome (Lynch *et al.*, 1966) causes a predisposition to colorectal cancer via germline pathogenic variants in mismatch repair genes (*MLH1, MSH2, MSH6, PMS2,* and *EPCAM).* Originally named hereditary non-polyposis colorectal cancer (HNPCC) to describe the primary associated phenotype, this syndrome was later renamed Lynch to account for the other cancer phenotypes that were also attributed to pathogenic variants in these genes. Lynch syndrome is currently also associated with an increased risk of uterine, stomach, breast, ovarian, and pancreatic cancer (Engel *et al.*, 2012).

The story of novel associations between cancer phenotypes and pathogenic variants in known cancer genes is being increasingly described in the literature. A recent study of germline variants in 10,389 adult cancers using The Cancer Genome Atlas (TCGA) cohort described an association between loss of function variants within homologous recombination gene *PALB2* and stomach adenocarcinomas (Huang *et al.*, 2018). A similar observation has previously been described by Sahasrabudhe et al, linking pathogenic germline variants in homologous recombination genes *PALB2, BRCA1,* and *RAD51C* with HDGC and sporadic gastric cancer (Sahasrabudhe *et al.*, 2017). Additionally, the TCGA cohort study found loss of function germline variants in *SDHA* in melanoma, despite these being predominantly associated with predisposition to paragangliomas and pheochromocytomas (Burnichon *et al.*, 2010). This adds complications to the story of cancer predisposition. Particularly in familial cases showing generations of unusual genotype-phenotype associations, it appears as though other currently unknown genetic factors influence the presenting phenotype.

The concept of cancer genes causing variable phenotypes has been previously explored by Whitworth et al (Whitworth, Skytte, Sunde, Derek H. Lim, *et al.*, 2016). The study designated the term Multilocus Inherited Neoplasia Alleles Syndrome (MINAS) to define cases with multiple pathogenic variants in cancer associated genes, and suggests that these may produce extreme or unexpected phenotypes (Whitworth, Skytte, Sunde, Derek H. Lim, *et al.*, 2016). An example of this is the extreme phenotype of macrocephaly, multifocal papillary thyroid cancer, paraganglioma of the left common carotid artery, and paraganglioma of the right carotid body (all diagnosed within a three year time period) in an individual with germline variants in *PTEN* and *SDHC* (Zbuk *et al.*, 2007). The study also demonstrated a number of cases where the phenotype is not associated with pathogenic variants in either of the identified genes, such as pathogenic variants in *BRCA1* and *PALB2* in an individual diagnosed with uterine myomas and a meningioma prior to a diagnosis of breast invasive ductal carcinoma which is more typical of the identified genotype (Pern *et al.*, 2012).

Within most cancer phenotypes, there remains some degree of missing heritability. Within breast cancer, around 50% of familial risk remains unexplained (Skol, Sasaki and Onel, 2016). Genome wide

association (GWA) studies have played a great part in uncovering breast cancer associated variants, with over 80 loci being identified which explain around 18% of heritability (Michailidou *et al.*, 2015, 2017; Skol, Sasaki and Onel, 2016). According to the 'common disease, common variant' hypothesis (Pritchard and Cox, 2002), heritability of rarer cancers such as gastric cancer, and in particular the diffuse subtype, are less likely to be explained by GWA studies which typically look for variants that have an allele frequency of above 1% in the tested population. One GWA study identified a particular association between gastric cancer and loss of function variants in DNA repair gene *ATM*, however this study used a homogenous Icelandic population and so was possibly subject to a founder effect, making the identification of rare variants possible in this case (Helgason *et al.*, 2015).

The rarity of certain cancers also greatly limits the ability for large sequencing studies to be performed due to limited participant numbers. In these cases, utilising familial data and populations that are enriched for genetic disease predisposition can allow for the identification of rare disease associated variants in small, statistically underpowered datasets. Recruiting and sequencing affected and unaffected individuals from families allows researchers to select variants that co-segregate with phenotype using a high to moderate penetrance, autosomal dominant inheritance model. This technique has been successfully utilised to further explore the role of pathogenic variants in *PALB2, BRCA1,* and *RAD51C* in diffuse gastric cancer families (Sahasrabudhe *et al.*, 2017).

Homogenous populations and those enriched for disease have also been key to identify rare disease predisposing variants. Disease predisposing founder variants are often discovered in previously geographically isolated regions (e.g. Quebec and Newfoundland), countries (e.g. Finland and Iceland), and ethnic groups (e.g. Ashkenazi Jewish populations) where a limited population size leads to low levels of genetic variation (Ponti *et al.*, 2015). Around 55 founder variants have been discovered in Lynch syndrome, proving clinically useful for panel testing of colorectal cancer families (Ponti *et al.*, 2015). In particular, sequencing studies within the Finnish population identified two pathogenic variants in *MLH1* which are responsible for around 50% of Finnish Lynch syndrome cases (Nyström-Lahti *et al.*, 1995; Moisio *et al.*, 1996; Ponti *et al.*, 2015).

The basic concept of selecting study participants who are more likely to be affected by genetic predisposition factors can be achieved in several different ways. On one hand, one can select study participants who are not nor have ever been exposed to environmental risk factors; for example, studying lung cancer risk in individuals who have never smoked (Lan *et al.*, 2012). However this removes the ability to study the combinatory effect of genetic and environmental risk factors as was done in a GWA study of smokers and non-smokers with lung cancer by McKay et al (McKay *et al.*, 2017). The selection of young onset cases also enriches the case population for genetic factors, as demonstrated by Wasserman et al who showed that young onset adrenocortical carcinoma (ACC) cases are more likely to carry germline, pathogenic *TP53* variants (Wasserman *et al.*, 2015). Some studies

have utilised the identification of multiple primary tumours as an indication of increased risk of germline cancer predisposing factors (Cybulski, Nazarali and Narod, 2014). One study that used this model is the Women's Environment, Cancer and Radiation Epidemiology (WECARE) GWA study of individuals with contralateral in comparison to unilateral breast cancer, identifying single nucleotide polymorphisms (SNPs) that modify contralateral breast cancer risk (Teraoka *et al.*, 2011).

These techniques for study design provide researchers with the ability to study rare disease-causing variants in smaller populations which are often limited in the study of rare diseases. In an age of data sharing, researchers are fortunate to also have access to external resources such as the TCGA cohort. Similarly, sequencing studies are encouraged to make data publicly available, providing the opportunity to collect and collate sequencing data from a variety of different resources. With some rare cancers such as ACC and diffuse gastric cancer, public datasets are still limited in size. It is therefore important that studies such as those described in this thesis provide new sequencing data to the scientific community to increase understanding of predisposition to rare cancers and diseases.

1.2 The process of oncogenesis

The hallmarks of cancer describe a set of capabilities acquired by a healthy cell as it undergoes oncogenic transformation. In 2000, Hanahan and Weinberg described six capabilities of cancer cells, including sustaining proliferative signalling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, and resisting cell death (Hanahan and Weinberg, 2000). These capabilities arise in part through variants in key regulatory genes. It is estimated that a human cell acquires around 70,000 genomic lesions per day, of which around 75% are single-stranded DNA breaks (Lindahl and Barnes, 2000; Tubbs and Nussenzweig, 2017). In healthy cells the DNA undergoes extensive surveillance and repair of damage, preventing a large-scale accumulation of DNA damage. Additionally, a complex checkpoint system forces damaged cells into either senescence or apoptosis (Jackson and Bartek, 2009). Often cancer cells evade these two genomic maintenance systems, creating a feedback loop of further damage and pathway evasion.

The process of key pathway evasion can begin at birth for carriers of cancer predisposing variants. Many typical cancer genes, including *BRCA1, BRCA2, TP53,* and *PALB2* function in the DNA repair pathway; it has been shown that a large proportion of breast cancer patients with germline pathogenic *BRCA1* or *BRCA2* variants lose the second copy of the gene in the tumour (Staff *et al.*, 2000; Osorio *et al.*, 2002; Simon and Zhang, 2008). This DNA repair deficiency can produce a "mutator phenotype", creating genomic instability that facilitates further somatic variants (Loeb, 2001; Chae *et al.*, 2016). This mutator phenotype can also be caused by somatic driver mutations in DNA polymerase ε and δ, producing an ultra-hypermutated cancer which has been described in the context of mismatch repair deficient childhood brain tumours (Shlien *et al.*, 2015). Additionally, germline variants in the proofreading domains of these two polymerases have been shown to predispose to colorectal cancer, with 16% of colorectal tumours being hypermutated in the TCGA database (Muzny *et al.*, 2012; Loeb, 2016). An array of different variants can be produced in this manner, the majority of which are dubbed 'passenger' lesions, which do not contribute to a malignant or invasive phenotype (Pon and Marra, 2015). However some pathogenic variants result in the activation of oncogenes, or inactivation of tumour suppressor genes, giving cells a particular selective advantage in the tumour microenvironment (Stratton, Campbell and Futreal, 2009; Pon and Marra, 2015).

The idea of cancer as an evolutionary process was described by Peter Nowell in 1976 (Nowell, 1976), suggesting that somatic variants are selected for in a tumour microenvironment in a mechanism mirroring Darwinian natural selection. The tumour micro-environment provides selective pressures, within which cells compete for space and resources. The majority of cells don't succeed, as demonstrated by the tumour doubling time which has been shown to be markedly slower than the cellular doubling time, which suggests that the majority of cancer cells are unable to survive to mitosis (Klein, 2009).

Those cells which do survive have often gained the ability of 'sustained proliferative signalling', the first of Hanahan and Weinberg's hallmarks of cancer (Hanahan and Weinberg, 2011). An example of this can be seen in melanomas, of which a study showed 42% of tumours carried somatic pathogenic *BRAF* variants which caused constitutive signalling of the mitogen activated protein-kinase (MAPK) pathway (Hocker and Tsao, 2007; Davies and Samuels, 2010). In additional studies, 73% of melanoma samples have been shown to carry pathogenic variants in 13 different genes which activate the MAPK and PI3K signalling pathways (Shain *et al.*, 2015). The MAPK pathway is a key regulator of proliferation and differentiation in melanocytes and is hyperactivated by UV damage induced receptor tyrosine kinase stimulation or by pathogenic variants, allowing for uncontrolled melanocyte proliferation (Wellbrock, 2014).

In addition to sustaining proliferation, cancer cells must be able to evade growth suppressors. One of the most well studied tumour suppressor genes is *TP53*, which has been shown to be one of the most highly mutated genes in tumours, being mutated in 26.9% of all cancers from the catalogue of somatic mutations in cancer (COSMIC) database (Chae *et al.*, 2016). Inactivation of tumour suppressors is often caused by loss of function variants, however the majority of *TP53* somatic variants (73.4%) are missenses which can cause a gain or loss of protein function (Olivier, Hollstein and Hainaut, 2010). The proportion of missense variants in *TP53* that are predicted to create a non-functioning protein, as measured by the transactivation activity of the protein generated by single nucleotide variants, changes by cancer type. Transactivation activity was measured in mutated proteins in comparison to the wild-type protein and variants were categorised as 'supertrans', 'functional', 'partially functional' or 'non-functional' (Petitjean *et al.*, 2007). In colorectal cancer, 84.9% of somatic pathogenic *TP53* variants were predicted to create a non-functioning protein, this is in comparison to 66.7% of skin squamous cell carcinomas (SCC) tested (Petitjean *et al.*, 2007). The tumour suppressor p53 acts in a pathway receiving intracellular signals such as genome damage, nucleotide pool levels, and growth-promotor signals and can halt the cell-cycle progression accordingly (Olivier, Hollstein and Hainaut, 2010; Hanahan and Weinberg, 2011).

The apoptotic machinery receive both extrinsic signals involving tumour necrosis factors on the cell surface, and intrinsic signals regulated by the BCL2 protein pathway (Adams and Cory, 2007). These signals go on to activate a caspase pathway which executes cell apoptosis. One of the sensors involved in detecting DNA damage uses the p53 tumour suppressor pathway. In response to DNA damage and chromosomal abnormalities, p53 removes cells from the cell cycle and signals for cell death. Therefore, variants in *TP53* as previously described can help cells evade apoptosis. In addition to loss of p53, BCL2 has been described as influencing tumorigenesis. Somatic pathogenic variants in the promoter region have been associated with increased gene expression and apoptosis evasion in breast cancer cells (bhushann Meka *et al.*, 2016). Additionally expression of BCL2 has been linked to chemotherapy and

radiotherapy response, making it an interesting prognostic factor for breast cancer patients (bhushann Meka *et al.*, 2016).

Another related tumour mechanism is inducing replicative immortality. Healthy cells are limited in their potential to replicate by eroding telomeric regions at the end of chromosomes. In immortalised cells, a specialised DNA polymerase, Telomerase, adds telomeric repeat sequences to chromosomal ends preventing erosion and causing a resistance to senescence. Somatic pathogenic variants in the promoter of telomerase reverse transcriptase gene (*TERT*), leading to a promotion of expression, have been described in many cancer types, and have been shown to be linked to outcomes and prognosis in bladder cancer (Li *et al.*, 2015).

Angiogenesis is the process of vasculature development that provides tumours with nutrients and oxygen and removes metabolic waste (Hanahan and Weinberg, 2011). In adults, angiogenesis is only switched on temporarily in the event of wound healing or female reproductive cycling. However, this process is often switched on during tumour progression due to overexpression of regulators such as vascular endothelial growth factor-A (VEGF-A). Somatic amplification of this gene has been seen in a number of tumour types, in particular in hepatocellular, pancreatic and intestinal adenocarcinomas, and large cell carcinomas of the lung (Andreozzi *et al.*, 2014).

The development of an invasive or metastatic tumour phenotype is associated with late stage and aggressive disease. To facilitate this, cancer cells often change shape and detach from other cells and from the extracellular matrix. One key protein involved in the adhesion to the extracellular matrix is E-cadherin, encoded by the gene *CDH1*. Reduced expression of this gene somatically and its adhesion partner beta catenin (*CTNNB1*) has been linked to poor prognosis in oesophageal cancer (Ishiguro *et al.*, 2016), in addition to reduced expression in gastric and breast cancer (Asiaf *et al.*, 2014; Hansford *et al.*, 2015).

The expression of genes in proliferation, angiogenesis and invasion-related pathways has been shown to be a key process in oncogenesis. However, a number of these pathways are also affected by germline cancer predisposing variants. The initiation of genomic instability by many cancer risk genes has previously been discussed. However other cancer predisposition genes have been described that target different cancer hallmarks. Von Hippel-Lindau disease is caused by pathogenic variants in the *VHL* gene and is associated with various tumours including clear-cell renal cell carcinoma and phaeochromocytomas (Gossage, Eisen and Maher, 2015). Loss of the second wild type allele of *VHL* has been associated with overexpression of growth factor VEGF inducing angiogenesis in a HIF-α dependent manner (Gnarra *et al.*, 1996; Gossage, Eisen and Maher, 2015). Germline pathogenic variants in extracellular matrix genes *CDH1* and *CTNNA1* are associated with an increased risk of gastric cancer (Pharoah, Guilford and Caldas, 2001; Slavin *et al.*, 2017). However, the majority of cancer cells rely on the occurrence of genome instability as an early event to provide genotypic variability allowing for

the selection of further favourable variants. This can occur through classic DNA repair deficiency variants, exposure to UV-damage, or in the case of pancreatic cancer, through the simultaneous development of genome rearrangements, possibly facilitated through replication or transcriptional stress (Notta *et al.*, 2016; Tubbs and Nussenzweig, 2017).

1.3 Whole exome sequencing

The human genome consists of around 3 x 10^9 bases, of which only 1% code for proteins. According to data from the 1000 genomes project, the average human differs from the reference (GRCh37) at 4.1 to 5.0 million sites (Auton *et al.*, 2015). The vast majority of these variants are within the 99% of the genome which is not protein coding. Of the small number of exonic (protein coding) variants, around 11,000 contain synonymous variants which cause no change in amino acid (Auton *et al.*, 2015). An estimated 10,000 to 12,000 variants alter or truncate the protein sequence (Auton *et al.*, 2015). The majority of germline variants found naturally among populations are not disease causing. However it has been estimated that 85% of disease causing variants are within protein coding or regulatory regions (Rabbani, Tekin and Mahdieh, 2014). For this reason, when looking for disease causing variants, a whole exome sequencing (WES) strategy is often employed to only identify those variants in protein coding regions.

The first sequencing technologies focused on simple RNA species such as tRNA and single-stranded RNA bacteriophages. In 1965, two separate teams led by Robert Holley and Fred Sanger developed related techniques to sequence RNA fragments (Holley *et al.*, 1965; Sanger, Brownlee and Barrell, 1965). The technique developed by Sanger used two-dimensional fractionation and radiolabelling to detect partial-digestion fragments (Sanger, Brownlee and Barrell, 1965). This 2-D fractionation technique was used by Walter Fiers in 1972 to produce the first protein coding gene sequence, followed by the complete genome sequence of bacteriophage MS2 (Jou *et al.*, 1972; Fiers *et al.*, 1976). After a number of rapid developments in the field of nucleotide sequencing, in 1977 Sanger developed the 'chain-termination' technique, using deoxyribonucleotides (dNTPs) to produce sequences of increasing length and label the final nucleotide for detection and therefore infer the targeted DNA sequence (Sanger, Nicklen and Coulson, 1977). This technique became known as Sanger sequencing and formed the basis for first-generation sequencing machines.

Second-generation sequencing techniques took a different approach, measuring luminescence produced by pyrophosphate conversion to ATP, and subsequently to luciferase, as nucleotides are washed sequentially over template DNA (Nyrén and Lundin, 1985). This pyrosequencing technique was licensed by biotechnology company 454 Life Sciences where it developed into commercially successful next-generation sequencing (NGS) technology (Heather and Chain, 2016). Technological developments including the generation of flowcells to bind DNA strands allowed for the creation of paired-end sequencing, usually of short reads, providing greater accuracy when mapping to reference sequences. Further advances in technology gradually led to the development of longer read sequencing techniques, and importantly, lower costs at a greater read depth, making the technology viable in a research and clinical environment.

A WES strategy has since been applied in many different research contexts, providing high depth, protein-coding data that is cheap enough to make large scale projects financially viable. In addition to its lower costs when compared to whole genome sequencing (WGS), aligned and merged WES data are more compact. An average aligned whole exome BAM file (at 60x coverage) will take around 4GB of disk space, in comparison to a whole genome file (at 60x coverage) which can take around 150GB. This vastly different file size means that one has to factor in the costs of greater data storage for WGS projects, in addition to greater computational power for alignment. These factors have made WES more practical in a clinical and diagnostics setting, where potential candidate genes may not be covered by a targeted panel, or for heterogenous diseases such cardiac diseases and intellectual disabilities (Y. Sun *et al.*, 2015).

The advent of GWA studies has proved invaluable in identifying common SNPs that contribute to complex traits such as autoimmune diseases and cancer and allows for the interrogation of lesser studied regulatory regions (Michailidou *et al.*, 2015; Ji *et al.*, 2016). This approach utilises the principle of linkage disequilibrium (LD; the non-random association between alleles positioned at different locations) to map disease causing variants to marker alleles which are genotyped on an array. Using this principle, one can impute genotypes and test for associations for over 10 million common SNPs in non-African populations by genotyping only 500,000 (Belmont *et al.*, 2005). Some have pointed out that GWA studies have not lived up to the high expectations that they will explain large proportions of complex trait heritability. An example of this is the studies of human height, which has an estimated heritability of around 80%, of which three research consortia have reported 54 phenotype associated variants (Visscher, 2008). However the variants identified in this study of ~63,000 people explain only 5% of phenotypic variance (Visscher, 2008; Manolio *et al.*, 2009). Despite many different GWA studies covering hundreds of thousands of individuals, a degree of missing heritability still exists. This has led some to suggest that the common disease-common variant hypothesis has been found not to apply to the majority of complex human diseases (Visscher, 2008; Manolio *et al.*, 2009; Visscher *et al.*, 2012).

As sequencing costs continue to drop, it is now feasible to perform genome wide studies on WGS data. A number of studies make use of lower coverage WGS to detect rare variants or structural variants, then impute these into pre-existing GWA study data (Gudbjartsson *et al.*, 2015; Luo *et al.*, 2017). This opens new avenues for exploring rare variants in large cohorts and may provide a balance between rare variant detection and the affordability of genotyping.

1.4 Development of sequencing analysis techniques

There is no single approach to analysing WES data, although best practice guides do exist to guide researchers through the basics of alignment, processing, variant calling, hard filtering, and variant annotation. The Genome Analysis Toolkit (GATK) provides a number of tools for use in quality control, data manipulation, and variant discovery (Auwera *et al.*, 2014). Although the basic sequencing analysis processes largely remain the same between pipelines, which tools are used to undertake these processes can vary greatly depending on ease of installation, personal and institutional preference, and concordance with previous work.

In addition to differences in sequencing pipeline composition, researchers can apply different technical filters to their sets with the aim of improving overall variant quality. Next generation sequencing technologies are associated with greater error rates in comparison to the more traditional capillary based methods (Koboldt *et al.*, 2010). To account for this, often filters are applied in addition to the recommended GATK Variant Quality Score Recalibration (VQSR) filtering step. The GATK VQSR filtering creates an adaptive error model based on variant quality by depth, mapping quality, variant position within reads, and strand bias of validated variants from the HapMap project, then uses this model to calculate the probability that variants are real (Carson *et al.*, 2014). GATK recommends a VQSR threshold of 99% sensitivity for true variants, however low quality genotypes and unvalidated variants have been shown to pass these filters (O'Rawe *et al.*, 2013).

Genotyping errors can cause problems for association testing, particularly when they differentially affect cases and controls, and have been shown to negatively impact both common variant and aggregated rare variant association tests (Mayer-Jochimsen, Fast and Tintle, 2013). Although non-differential errors which occur equally in cases and controls do not affect type I error rate, they have been shown to significantly decrease statistical power, particularly in rare variant association tests (Powers, Gopalakrishnan and Tintle, 2011). Importantly, it is noted by GATK that VQSR does not account for low quality genotypes which are a major source of sequencing error, and so recommends additional downstream genotype and dataset specific filters (McKenna *et al.*, 2010). One study that tested the use of additional filters recommended genotype filtering on read depth (DP) and genotype quality (GQ) metrics generated by the GATK variant caller (Carson *et al.*, 2014). This study also suggested that after genotype filtering, implementing a call rate filter provided the greatest data quality improvement, as calculated by testing the transition/transversion (Ti/Tv) ratio before and after filters (Carson *et al.*, 2014).

Once detected and filtered, the discovery of associations between rare variants and disease can also provide difficulties. As previously discussed, gathering rare disease cases in great enough numbers for the detection of rare, disease associated variants is not always feasible. As such, association tests can be statistically underpowered and provide inaccurate estimates of rare variant effect (Hoffmann, Marini

and Witte, 2010). One technique to increase statistical power for small sample sizes is to aggregate rare variants into functional groups. This can be done by summing variant counts in cases and controls. A comparable technique was performed by Morgenthaler et al to identify variant carrying genes that confer a risk to disease and was defined as a cohort allelic sums test (CAST) (Morgenthaler and Thilly, 2007). However this test does not weight variants and therefore assumes that all tested variants carry the same direction of effect (Morgenthaler and Thilly, 2007).

A similar variant aggregation approach can be taken that includes variant weighting, with the primary aim of upweighting variants that are most likely to be disease causing. The Madsen and Browning approach weights alleles by the inverse of the estimated standard deviation of the total number of variant alleles in unaffected samples (Madsen and Browning, 2009; Hoffmann, Marini and Witte, 2010). Variants can be gathered into functional regions such as genes, that region is then tested for an association. In comparison, the Combined Multivariate and Collapsing (CMC) method aggregates rare variants into functional regions but analyses more common variants on an individual basis (Li and Leal, 2008). This analysis provides the benefit of rare variant aggregation without reducing the power for detecting individual common variant associations.

Outside the realms of variant association tests, there are many tools available for variant prioritisation and candidate selection. These tools may be less necessary for familial studies where segregation analysis can provide a basis for highlighting candidate variants. However, where datasets consist of affected, unrelated individuals, these tools can provide the basis to only study variants that are most likely to cause disease predisposition. In smaller datasets, filtering variants based on the frequency of that variant within the case set is less informative. In these sets one might utilise external control cohorts such as the 1000 genomes project data to exclude common variants (Auton *et al.*, 2015).

The third phase of the 1000 genomes project describes the genomes of 2,504 individuals from 26 populations (Auton *et al.*, 2015). The study utilises low-coverage WGS, high coverage WES, and microarray genotyping to describe over 88 million SNPs, short insertions and deletions, and structural variants (Auton *et al.*, 2015). Study participants are divided into five ethnic 'super populations' which can be used to ascertain specific population based allelic frequencies. Similar control sequencing cohorts such as the Exome Aggregation Consortium (ExAC) data also provide population specific allele frequencies (Lek *et al.*, 2016). The ExAC dataset includes WES data of 60,706 individuals gathered from large sequencing consortiums, some of which are disease affected such as the Myocardial Infarction Genetics Consortium and TCGA (Lek *et al.*, 2016). The latest version of ExAC, renamed Genome Aggregation Database (gnomAD) to reflect the inclusion of WGS data, is larger than its predecessor, containing 123,136 exomes and 15,496 genomes. However, unlike ExAC, a version of gnomAD has not been released which excludes TCGA data and so is it is currently not a valid control set for cancer association studies.

Other control sets that could be applicable include the 1958 Birth Cohort project and subsequent ICR1000 project which provides users with access to germline WES data from 1000 white British individuals born in one week in 1958 (Ruark *et al.*, 2015). Although not publicly available, this set provides individual exome sequences from an ethnically controlled population and so could be a useful control set to researchers studying disease association in British individuals (Ruark *et al.*, 2015).

In addition to the exclusion of variants that commonly appear in healthy individuals, variant consequences can be predicted and used to select those that are most likely to be protein-affecting. The variant effect predictor (VEP) tool by GATK assigns a variant consequence based on sequence information in the predicted canonical transcript or in all transcripts (Auwera *et al.*, 2014). The VEP tool also applies other well-used variant prediction algorithms to predict whether variants might affect protein structure or lie within key binding regions and therefore have a deleterious effect on protein function. This is less relevant for variants that are predicted to cause loss of protein function due to an early stop codon or shift in reading frame. However, predictions are important (and in some cases only applied) to variants that cause a one amino acid substitution, allowing users to prioritise variants by effect.

Methodologies for variant prediction tools can be divided into four approach styles: sequence conservation, protein structure, combined structural and sequence approach, and meta-predictors that integrate results of other prediction tools (Tang and Thomas, 2016). By studying sequence homology, tools such as the 'Sorting Tolerant from Intolerant' (SIFT) algorithm predict the effect of all possible substitutions at given amino acid positions (Kumar, Henikoff and Ng, 2009). This is based upon the assumption that key amino acids in a protein are conserved throughout evolution. The same principle is used by the MutationAssessor functional impact score to estimate variant pathogenicity (Reva, Antipin and Sander, 2011). In contrast, machine learning algorithms such as AUTO-MUTE use databases of known protein structures to calculate how amino acid changes affect protein structure stability (Masso and Vaisman, 2010).

PolyPhen-2 is a widely used variant prediction tool that combines information about sequence conservation with knowledge of structural protein-affecting factors (Adzhubei, Jordan and Sunyaev, 2015). The tool feeds information from sequence, phylogenetic, and structural factors into a classifier trained on known pathogenic and non-pathogenic variants to predict the probability that the queried variant is damaging (Adzhubei, Jordan and Sunyaev, 2015). SIFT and PolyPhen-2 are some of the most well used prediction techniques and multiple scores have been developed integrating the two approaches. CAROL is one such scoring system, using a linear weighted Z-score to combine SIFT and PolyPhen-2 (Lopes *et al.*, 2012).

Which tools a researcher implements and how they select cut-offs for variant filtering should be decided on a per dataset basis. For example, stringent variant rarity filters using control sets will be more

appropriate for the study of extremely rare disease predisposition, or where complete penetrance is expected. Many studies have been conducted comparing sequencing pipelines (Pabinger *et al.*, 2014), variant filtering strategies (Carson *et al.*, 2014), and available prioritisation tools (Tang and Thomas, 2016) to help researchers choose the most appropriate methods for their data.

1.5 Aims

The purpose of the work presented in this thesis is to answer questions about the missing heritability in cancer syndromes by studying the whole exome sequencing data from young onset cases, familial cases, and multiple primary tumour cases using novel variant analysis techniques. The aims are as follows:

1. To identify novel associations between genes with protein-affecting germline variants and a predisposition to hereditary diffuse gastric cancer in families without pathogenic *CDH1* variants.
2. To identify if genetic factors cause the development of sweat duct proliferations and an irregular distribution of elastin fibres in the skin.
3. To identify predisposing genetic factors associated with adrenocortical carcinoma development with a particular focus on the oncocytic subtype.
4. To study predisposition genes in unilateral and contralateral breast cancer cases from two research cohorts.
5. To create novel data analysis approaches to identify genetic factors associated with rare disease predisposition and development.

Although this thesis deals more generally with cancer predisposition and the issues surrounding whole exome sequencing analysis for rare diseases, it will also describe each studied phenotype in more detail within chapters. Accordingly, information about known predisposing factors, pathology, and cohort specific methods can be found within each chapter, while this introduction and the methods chapter have been aimed towards describing overarching concepts, themes, and analysis methods.

**Germline WES pipeline was created by bioinformatician Alexey Larionov, and updated and maintained by Alexey and me*

2 Materials and Methods

The number and variety of different sequencing datasets explored within this study required a range of different analysis techniques. After sequencing, many of the sets required the same germline WES VCF generation pipeline, generated in-house*. However somatic and RNA sequencing pipelines were generated specifically for some studies presented in this thesis. Additionally, tools were developed to create more efficient or thorough analysis workflows as required, and where possible were made available for other studies. This included tools that provide unique information for variant filtering and candidate selection. During the process of tool creation, temporary solutions were often created prior to the generation of the final tool version. These will not be discussed in this section, which focuses on complete pipelines and tools that were developed for use in this study. All tools and pipelines discussed here are freely available on GitHub (https://github.com/elliefewings).

2.1 Germline whole exome sequencing VCF generation pipeline*

The eight-step germline WES VCF generation pipeline was run on the High Performance Computing (HPC) Cambridge Service for Data Driven Discovery (CSD3) platform. The computing clusters required the use of Simple Linux Utility for Resource Management (SLURM) systems to submit and queue jobs. Therefore, each stage of the pipeline required a multistep job submission process of data copying, job performance, and summarising prior to copying data back to the long-term storage blades.

2.1.1 Alignment and quality control

The raw basecall (BCL) files generated by Illumina sequencing instruments were converted to FASTQ files using the Illumina bcl2fastq tool. During this processing step, completed by the sequencing facility, demultiplexing of multi-sample sequencing lanes and trimming of adapters was performed. This study generated paired-end sequencing data and so prior to alignment a sample file (giving the sample name, first FASTQ, second FASTQ and md5 checksums file) was generated. Prepared data for each lane of sequencing were passed to the alignment script in a job description file, specifying data locations on storage blades and tool versions amongst other parameters required for the script.

FastQC reports were generated for raw FASTQ files. The Burrows-Wheeler Aligner (BWA)-MEM (Li and Durbin, 2010) algorithm (v.0.7.12) was used for efficient long read alignment to hg19. Samtools (Li *et al.*, 2009) (v1.2) fixmate fills in mate coordinates of sorted BAM files prior to cleaning with Picard (http://broadinstitute.github.io/picard) (v1.133) CleanSam and the addition of read groups with the AddOrReplaceReadGroups option. Summary metrics were created with the generated files using Flagstat by Samtools and Qualimap (Okonechnikov, Conesa and García-Alcalde, 2015) (v2.1.1).

These alignment steps were performed per lane of sequencing prior to merging.

2.1.2 Lane Merging

Often samples are run across several lanes of sequencing for increased read depth and quality. However, if a single lane of sequencing is given to this stage of the pipeline, merging is skipped. If multiple lanes of sequencing data were available, Samtools was used to merge the BAMs per sample across each lane of sequencing. Picard MarkDuplicates was used to mark PCR duplicates which were later removed using Samtools view. The deduplicated BAM files were indexed and md5 sums regenerated prior to further quality control, including Picard's CollectInsertsSizeMetrics and CollectAlignmentSummaryMetrics, and other metrics generated by Qualimap and Samstat (Lassmann, Hayashizaki and Daub, 2011) (v.1.5.1).

2.1.3 BAM pre-processing and gVCF generation

Once BAMs from each library had been merged across lanes, they underwent local realignment around insertions and deletions with GATK (McKenna *et al.*, 2010) (v3.6.0) IndelRealigner tool. All GATK tools allow the user to set padding around target intervals; a padding of 10 base pairs around exons was introduced at this stage. BAM files were checked for any additional PCR duplicates that had been created by realignment, these were again marked and removed. Files underwent base quality score recalibration (BQSR) with GATK using a generated BQSR table to detect sequencing errors by estimating base call quality scores. The processed BAMs underwent variant calling using GATK's HaplotypeCaller in gVCF mode to generate a gVCF file per sample.

2.1.4 Combining gVCFs

At this stage, sample gVCFs were combined into datasets for joint variant calling. These datasets are often of one phenotype, so analysis was completed generating one gVCF per phenotypic group. For each dataset, gVCFs were copied from their libraries into a common location and ran through GATK's CombineGVCFs. At this stage, our standard intervals padding of 10bp around exons were removed as this generates non-specific errors further downstream. One multi-sample gVCF with index and md5 was generated for further analysis.

2.1.5 Joint variant calling

Variant calling was performed on the combined gVCF. According to GATK best practices (Auwera *et al.*, 2014; Poplin *et al.*, 2017), this approach uses shared information across samples to provide greater sensitivity across areas of lower coverage. It also provides genotypes in all samples if one sample is called with a variant, allowing the data user to distinguish between reference sites and missing data.

GATK's GenotypeGVCFs was used to perform joint variant calling (using parameters: maxAltAlleles=6, stand_call_conf=30, stand_emit_conf=30) prior to filtering to remove variants not detected in any genotype. GATK's SelectVariants and VariantAnnotator were used to flag multiallelic variants. To improve data quality, GATK's VariantRecalibrator trains a variant quality score recalibration (VQSR) SNPmodel which was applied with ApplyRecalibration, using a SNP truth

sensitivity (TS) of between 97-99.5 depending on the set. For the HDGC set, this value was raised to 99.5 as it was noticed that previously validated variants were being filtered out using the standard 97 TS. A comparable indel model was trained and applied using the same tools, with a lower TS threshold of 95-97, again the higher end of this threshold was selected for the HDGC set. A histogram report was generated using a custom R markdown script and a summary of statistics on the output VCF was generated using BCFtools (Li, 2011) (v.1.2).

2.1.6 Hard filtering

Filters were selected based on VCF summary statistics generated after joint variant calling. Variants with a total depth across samples of less than 10 x the number of samples in the set (equivalent to an average depth of 10 per sample for each variant) were flagged. A QUAL filter was selected corresponding to a transition/transversion (Ti/Tv) ratio of 2, and those with score lower than this were flagged. These flagged variants were then removed using the GATK SelectVariants --excludeFiltered option.

2.1.7 Annotation

Filtered VCFs underwent splitting of flagged multiallelic variants using the GATK LeftAlignAndTrimVariants tool with the --splitMultiallelics flag. This ensured that multiallelic variants are treated as separate variants, and therefore are placed in separate rows in the VCF. Without this step, multiallelic variants can cause errors downstream as they often have several values, one for each alternative allele, in VCF fields. At this stage, variants were labelled with a unique ID. This ensured that the user could track variant information across different annotation files created further downstream. The GATK VariantAnnotator was used to add allele frequencies from the super-populations in the phase 3 1000 genomes control data (Auton *et al.*, 2015), and from the ExAC non-TCGA control data (Lek *et al.*, 2016). Variant Effect Predictor (VEP) (McLaren *et al.*, 2016) (v82) was used to annotate predicted variant consequences including gene names and SIFT (Kumar, Henikoff and Ng, 2009) and PolyPhen (Adzhubei, Jordan and Sunyaev, 2015) scores. The pick option was enabled to select the most likely canonical transcript for annotation.

2.1.8 Export

GATK VariantsToTable was used to export variants and genotype information into usable text formats for input into R and other downstream analysis. Each table was labelled with a code to describe the contents and unique variant IDs applied to each for indexing. A VEP table (VV) including predicted variant consequences was then exported. Separately, a table of previously applied 1000 genomes allele frequencies (kgen) and ExAC allele frequencies (exac) were exported for downstream use. Separate files were also generated for genotypes (GT), genotype qualities (GQ), sequencing depth (DP), allelic depths (AD), and phred-scaled likelihoods (PL scores).

These files were input into R scripts to check table dimensions (each file should contain information for the same number of variants and genotypes). In addition to the file containing standard alphabetic genotypes, these were converted to numeric codes and three additional tables are output based on the assumption of additive variant effect (homozygous reference =0, heterozygous=1, homozygous alternative= 2), dominant variant effect (homozygous reference =0, heterozygous=1, homozygous alternative= 1), and recessive variant effect (homozygous reference =0, heterozygous=0, homozygous alternative= 1).

The tables described above, including an md5 file, were output into text format for input into downstream genotype and variant filtering specific to each dataset.

2.2 Somatic whole exome sequencing VCF generation pipeline

A tumour specific pipeline was required using a somatic variant caller. Tumour data were called alongside data from a normal sample to call somatic variants that appear exclusively in the tumour sample. The pipeline uses similar processes to the germline pipeline, however omits certain incompatible stages such as gVCF generation. As the somatic variant caller used only calls one tumour/normal sample at a time, generation of gVCFs and joint variant calling is currently unsupported by variant calling tool MuTect2.

Additionally, somatic variant calling over just one sample takes a large amount of time, even when multithreaded. Each sample takes in excess of the maximum of 36 hours allowed for each requested node on the high-performance computers. Initially the variant calling step was run on an in-house server, with some whole exome tumour/normal pairs taking around 10 days for variant calling. To speed up this process, manual parallelization was completed by performing somatic variant calling in genomic chunks across different computing nodes on the high-performance computers. Generated VCFs for each genomic chunk were then combined into one sample VCF, which is combined into a multi-sample VCF if requested.

2.2.1 Alignment and quality control to lane merging

The first two scripts of the somatic pipeline used the same processes as the germline pipeline. The first included alignment of FASTQ files with BWA-MEM to hg19, filling in mate coordinates, and adding read groups before the generation of summary metrics as before. Lane merging was also comparable, using Samtools to merge and Picard to mark PCR duplicates before removal (see Germline whole exome sequencing VCF generation pipeline*: Alignment and quality control and Lane Merging)

2.2.2 BAM pre-processing

BAMs underwent the same pre-processing as in the germline WES pipeline, including local realignment, removal of introduced PCR duplicates, and base quality score recalibration. However variant calling was not performed at this stage. Processed BAM files for tumour and normal samples were output for further analysis.

2.2.3 Somatic variant calling

Apart from the absence of gVCF creation, the above stages of the somatic variant calling pipeline were comparable to the germline WES pipeline. At this point, BAM files from samples to undergo variant calling were collected into one location. This stage required a file containing the sample name, the name of the processed tumour BAM file, and normal BAM file for each sample. To multi-thread this analysis, the job submission script reads this file and requests 20 new computing nodes for each sample, running variant calling between the tumour and normal files over different intervals.

The Nextera Rapid Capture Exome (Illumina) intervals file was split into 20 chunks. Chunks 1 to 19 cover chromosomes 1 to 19, whereas chunk 20 incorporates chromosomes 20, 21, 22, X, Y, and mitochondrial DNA (MT). Collectively these last chromosomes are still comprised of fewer intervals than chromosome 1 alone, and so variant calling within chromosome 1 is likely to be the rate limiting factor within this analysis. During job submission, the node was allocated a chunk number and a corresponding intervals file to run variant calling across.

GATK MuTect2 (Cibulskis *et al.*, 2013) was used for variant calling because of its high sensitivity, particularly at lower variant allele frequencies (VAFs) (Cai *et al.*, 2016). For each sample and each chunk, MuTect2 was called using maxAltAlleles of 6, stand_call_conf of 30 and tumor_lod of 4. The tumour lod score was lowered from the default of 6.3 which was found to generate very few variant calls, possibly due to a low sequencing depth in the supplied tumour samples.

Once variant calling was complete, the script checked the number of completed chunk VCFs. If all chunks are complete, GATK CatVariants tool (org.broadinstitute.gatk.tools.CatVariants) was used to concatenate the completed raw VCFs. MuTect2 outputs genotypes with columns 'Tumor' and 'Normal' for each sample. These were renamed to include the sample name prior to creating a multi-sample VCF. Additionally, the QSS score (sum of base quality scores for each allele) was removed as it consisted of two values (one for each allele) which causes errors during VCF merging.

If multiple samples had been input into variant calling, the script checks if each requested sample has a complete, concatenated VCF file. If all samples are completed, these are then merged with BCFtools merge option, creating one VCF for the input set.

2.2.4 Hard filtering

The GATK SelectVariants tool was used to trim variants not detected in any genotype, although by the nature of the somatic variant calling pipeline, these sites should not exist. The VariantAnnotator tool flags multiallelic variants for splitting further downstream. Summary stats were generated using BCFtools both before and after removal of variants that did not pass technical filters using GATK SelectVariants.

2.2.5 Annotation

The annotation step again shares similarities to the germline pipeline. Multiallelic variants were split using GATK's LeftAlignAndTrimVariants tool with the --splitMultiallelics flag. Unique variant IDs were assigned, and variants were annotated with 1000 genomes and ExAC population allele frequencies. Variant Effect Predictor (VEP) (v82) was used to annotate predicted variant consequences using the pick option to select the most likely canonical transcript.

2.2.6 Export

VCF files called by MuTect2 have fewer technical fields, and so previously exported GQ and PL scores were unavailable to the somatic pipeline. A VEP table (VV) including predicted variant consequences was exported alongside the applied 1000 genomes allele frequencies (kgen) and ExAC allele frequencies (exac). Additional files were generated for genotypes (GT) (including numeric codes for predicted additive, dominant and recessive inheritance patterns), allelic depths (AD), and tumour and normal LOD scores for each sample (LOD). All generated tables were checked for correct dimensions with an R markdown script.

2.3 RNA sequencing and differential expression analysis pipeline

The RNA sequencing pipeline was generated to analyse raw FASTQs downloaded from the TCGA database (Zheng *et al.*, 2016). Paired-end FASTQ files often need renaming to include the sample name before being entered into analysis.

Each script takes the following three parameters: -d allows the user to set the name of the dataset, -i specifies the location of data to be input, -o sets the output location of the results. This pipeline was ran on the local medical genetics server, and so due to limits in computational resources, samples were analysed successively.

2.3.1 Alignment

Prior to alignment, this script created a working directory within the analysis folder and copied over the specified source FASTQs. Within this directory, a samples file was created, listing each sample name, and the names of the first and second FASTQs. For each sample within the list, FASTQC (v0.11.6) was run on both FASTQs and alignment was performed with TopHat (Kim *et al.*, 2013) (v2.1.1) which uses Bowtie2 (Langmead and Salzberg, 2012) (v2.3.1) to align to hg19. The no-coverage-search option was implemented to speed up alignment which omitted coverage searching, a feature that is recommended for splice junction mapping particularly in short single end read sequencing. The script generated a directory for each sample, containing the FASTQC report for each FASTQ and an 'accepted_hits' BAM file as well as summary files.

2.3.2 Processing

The aligned BAM files required processing before they undergo differential expression analysis. A new samples file was created containing a list of all folders (and therefore samples) within the alignment directory. The SAMtools (v1.6) sort option sorts the aligned BAM file and rmdup was used to remove PCR duplicates. A FASTQC report was generated for the new processed BAM file and Samtools flagstat is run on the pre-processed and processed BAM file.

2.3.3 Differential expression analysis

An additional, optional parameter was accepted for this script, allowing the user to check differential expression between all combinations of input samples, or to input a case control file (with the -c option) and test for differential expression between two groups of samples. Prior to this, Cufflinks (Trapnell *et al.*, 2013) (v2.2.1) was used to quantify RNA expression over genes using a GRCh37.p13 build GTF file downloaded from ensembl (release-75) (ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/). For each sample, fragments per kilobase of transcript per million mapped reads (FPKM) were counted over the genes described in the GTF file.

If a case control file is not supplied, the script ran differential expression across all samples using cuffdiff by Cufflinks across genes in the above mentioned GTF file. If a case control file is specified,

the samples were split into comma separated lists of cases and controls and supplied to cuffdiff as different replicates of the same sample. This technique allowed expression to be summarised and compared between the two groups across genes, coding sequences (CDS) and isoforms. The output gene expression txt file was then used to select significantly differentially expressed genes.

2.4 Control data browsers

When exploring candidate variants and genes in germline WES data, it is useful to know variant allele frequencies and how variable that gene is in a healthy population. Many control datasets such as ExAC offer a web browser to explore genes and variants, giving allele frequencies in different ethnic populations. However, the ExAC control dataset includes germline data from TCGA. A version of the ExAC VCF that excludes TCGA data is available to download and explore, however this requires a certain degree of computational knowledge. The online web browser of ExAC does not offer the option of searching for non-TCGA data only, and so for cancer-focused researchers, genetic counsellors and clinicians who are not comfortable working with a VCF this is a great hindrance.

One of the tools developed as part of this thesis was an R shiny (https://shiny.rstudio.com/) hosted web app for interactively searching and exploring ExAC non-TCGA, 1000 genomes and the 1958 birth cohort control sets (Auton *et al.*, 2015; Ruark *et al.*, 2015; Lek *et al.*, 2016). The shiny apps underwent several stages of correction and optimisation. The methods below describe the ExAC non-TCGA app script development and deployment, which was then applied to each of the three mentioned datasets. The apps are available at the following link: https://medgenbrowser.wordpress.com/.

2.4.1 Initial data filtering

A memory limit of 1GB was given to any shiny app deployed on the RStudio server. Loading large amounts of raw data from control datasets will force the app to surpass this limit and the app will crash, generating a memory error. For this reason, and for ease of viewing, raw datasets were initially filtered to cut out unnecessary data fields and select only protein-affecting variants.

A raw VEP annotated ExAC (release 1) non-TCGA VCF file was downloaded. GATK (v3.8.0) tools SelectVariants and VariantAnnotator and BCFtools (v1.3.1) were used to mask, flag and split multiallelic variants. Info field annotations labelling heterozygous counts were removed and VEP (v91) was used to reannotate the VCF using the 'pick' algorithm to select the transcript that is most likely to be functional. GATK VariantsToTable was used to select specific fields including variant location, ID, population allele counts, and VEP annotated consequence which was also split into separate fields.

The annotated and separated table was saved to a txt file for further filtering in RStudio. Protein-affecting variants were selected, including loss of function variants, inframe indels, and predicted deleterious and damaging missense variants (by SIFT and PolyPhen respectively). No allele frequency filters were applied to the set. A local version of the app, ran using a Windows batch file, was also available including all ExAC variants without consequence filtering, however this was too large to run on the shiny servers.

To further cut down the file size, columns were removed, including population allele frequencies which can easily be calculated by the user from allele counts and numbers, and variant quality metrics. The

app allowed users to apply their own variant filters, during which the complete files are copied within the instance memory prior to these filters being applied. Therefore, to prevent memory related crashes, the complete data input into app needs to be less than half of the 1GB limit, allowing it to be duplicated during processing. The generated dataframe was saved in packaged Rdata format which expanded to 410.9MB, carrying 1,275,972 variants.

2.4.2 RStudio Shiny App

With the aid of Shiny by RStudio app development guides (RStudio Inc, 2013), the described app was developed using RStudio (v0.99.903) and the Shiny library (v1.0.5). The app development process comprises of two parts, 1) the development of the user interface, 2) the defining of the server logic required to generate the desired output.

The user interface of the app utilised the fluidPage tool to assign outputs to certain sectors of the webpage. Using this tool, a title and description of the meaning of variant consequence filters is defined. Two input sections are specified, the first allows the user to input a string as a gene name and select whether that should completely or partially match the genes in the table, the second allows the user to select variant consequence (either loss of function only, or all protein-affecting variants (default)) and allele frequencies (< 0.05, < 0.01, or all (default)) to be included from dropdown menus. A download button was also created, which downloads the output file from the applied filters. The user interface also specifies a dataTable (DT library v0.4) output to display the filtered results.

The server section of the script takes the input filter values and uses a combination of reactive and eventReactive functions to filter the dataset. Reactives are implemented by the selection of an option from the dropdown menu, causing the dataset to be re-filtered every time a new option is chosen. EventReactives were used to filter on gene name when the user presses the 'search' button, this prevents the app from searching until the user finishes typing a gene of interest. Additionally, an option was supplied to allow the user to match the gene name exactly, changing the filtering type from a grepl function (base R), which matches partial strings, to the filter function in Dplyr (v0.7.4), which only selects complete matches.

In addition to variant filtering, when a gene is searched, summary metrics are created to describe occurrences of variants in that gene. A table was produced showing the total number of loss of function alleles and deleterious and damaging (by Sift and PolyPhen) missense alleles in each of the major ExAC populations (African, East Asian, European (Finnish), European (Non-Finnish), Latino, South Asian, and other) and a total across the set. This table is unaffected by consequence filters as it described both loss of functions and missenses. The counts however are affected by allele frequency filters and the filter selected is described at the top of the table.

The ggplot2 library (v2.2.1) was used to create a bar plot, summarising the allele frequencies of different consequence variants in the searched gene. The generated plot shows the numbers of loss of function and deleterious and damaging (by Sift and PolyPhen) missense alleles with an allele frequency below 0.01 and greater than or equal to 0.01 within that gene. This allowed the user to discern whether that gene contains many common or rare variants of that specific consequence category.

The apps for the ExAC and 1000 genomes are available to access via one web link, hosted by WordPress (https://medgenbrowser.wordpress.com/) which contains information about the apps and how to use them. As the 1958 cohort is not publicly accessible, this app is only available to run locally using either Windows batch script or a Mac command script depending on the users operating system. A full version of the ExAC non-TCGA data are also available in this format and can be requested using the contact form on the website.

2.5 Multigrep R package

The multigrep R package was originally developed as a function and placed into a package as it was frequently used in analysis. The premise of the function was to allow users to search for a vector of incomplete strings in a set. It was particularly used in this study to filter variants on consequence. As some variants are labelled with multiple consequence flags, one often has to filter by searching for any variant that contains the consequence of interest, and therefore is an incomplete match. This can be achieved with the grepl function, which takes one string and searches for incomplete matches across an object. However, grepl does not accept more than one string, and so this must be called multiple times when trying to search for several incomplete strings, for example when looking to find missense variants or stop gained variants. When searching for multiple complete matches, value matching can be achieved with the %in% operator (base R), which accepts a vector of values to match. The multigrep package aimed to combine both concepts and allow the user to search for a vector of strings for an incomplete match.

The multigrep function uses grepl inside a sapply (base R) across the vector of strings to match and returns a logical for each string in the object that was explored.

```
vect <- c("a", "b", "c")

x <- c("apple", "kiwi", "melon", "pear")

multigrep(vect, x)

[1] TRUE FALSE FALSE TRUE
```

2.6 Exon Variability Estimate (EVE) score

During analysis of rare variants across the different cohorts described in this work, several genes were consistently prioritised due to their carrying a number of different rare variants in a case set. Some of these genes, for example *TTN*, are large in size and so are likely to carry a large number of variants by chance. However, other genes were prioritised after more complex analysis, for example selecting genes that contain different rare variants across multiple affected families. Often this type of analysis removes genes such as *TTN* but prioritises others that contain rare variants that are unique to one family, despite variability in this gene being common. This is usually the case if a gene is functionally variable, examples of which include olfactory receptor genes or HLA genes whose variability is key to providing unique tastes or immune functions to individuals. These genes are often flagged as candidates in familial studies, where the gene contains multiple different rare variants in affected families. These genes can be removed using published lists of highly variable genes, such as those identified by the FLAGs study (Shyr *et al.*, 2014). However this study did not account for gene size when labelling those that are highly variable and so flagged genes like *TTN*, but not those that are smaller such as olfactory receptors (Shyr *et al.*, 2014).

2.6.1 Creating the EVE score

A VEP annotated VCF file from Phase 3 of the 1000 genomes project was used to generate a tool in R script which labels variants as within variable or invariable regions in a healthy population. Synonymous and non-synonymous variants from within this VCF file were organised into exons as determined by refGene coordinates (downloaded from UCSC table browser in 2016 with the following settings; clade: Mammal, genome:Human, assembly: Feb. 2009 (GRCh37/hg19), group: Genes and Gene Predictions, track: RefSeq Genes, table:refGene, output:GTF) and labelled with an exon number and the length of that exon in base pairs. Allele counts for each variant were determined from allelic frequencies for each subpopulation within the 1000 genomes project (AFR, AMR, EAS, SAS and EUR) as well as overall allelic frequencies. A variability score was generated by the number of alternative alleles per base pair in each exon, using a combined allele count of variants and the length of each exon. Labels were applied to each exon to show whether that score was in the lower quartile, mid-low quartile, mid-high quartile or upper quartile of all exons. A file was created of such scores and their corresponding exon positions that can be applied to any CSV format VCF within another R script.

To test the results of this scoring system, combined variability scores were generated for each gene by calculating the average variability scores across exons in a gene. Genes were sorted to determine the most variable within the 1000 genomes population and a word cloud was generated with the Tagxedo online tool (http://www.tagxedo.com/) to create results that were comparable to similar studies identifying variable genes (Shyr *et al.*, 2014) (figure 2.1).

Figure 2.1: Highly variable genes as determined by the average EVE score in each gene.

2.7 Sanger sequencing

Filtering of genomic data based on quality aims to reduce the number of false positives discovered, however inevitably there is always a risk of being too stringent and losing a large number of true positives. Filtering optimisation aims to find a balance between these, however validation of variants via Sanger sequencing is still required to confirm identified variants.

2.7.1 Primer design

To amplify candidate variants, flanking primers are designed. The Genome Compiler (http://www.genomecompiler.com/) software was used to provide a user-friendly, interactive design environment. Ensembl was used to select a large portion of the sequence surrounding the variant of interest, which was then copied into Genome compiler. Using the software's integrated IDT OligoAnalyzer (v3.1), forward and reverse primer sequences were selected that are between 17-30bp in length, and preferably at least 100bp up or downstream of the candidate variant. The total product size was less than 800bp.

Primers are designed with a melting temperature of around 60°C and with less than 2°C between primers. If possible, primers are designed with a 3' end that has multiple guanines or cytosines to improve binding. IDT OligoAnalyzer was used to predict secondary structures and homodimers that might occur. If a primer is predicted to homodimerize at greater than 4 bases, the primer was edited or redesigned to avoid this region.

Once the primer sequences were completed, they were checked using UCSC *in silico* PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr) to ensure that primers were specific to the region of interest. If multiple products were predicted the primers were redesigned.

2.7.2 Polymerase Chain Reaction (PCR) and gel electrophoresis

A PCR master mix was made using AmpliTaq Gold DNA polymerase with buffer I kit (Applied Biosystems). For each reaction, 2.5µl of 10x Reaction buffer (with $MgCl_2$), 0.5µl of dNTPs (10mM) (Invitrogen), and 0.125µl of AmpliTaq Gold DNA polymerase was made into a master mix and aliquoted out with 100ng of DNA, 0.5µl of the forward and 0.5µl of the reverse primer (10µM). The reaction was made up to 25µl with H_2O.

PCR reaction plates or tubes were placed on a BioRad Tetrad 2 (MJ Research DNA Engine Tetrad PTC-225) and ran for the following cycles; initial denaturing at 95°C for 10 minutes, 30 cycles of (denature at 95°C for 15 seconds, anneal at 60°C for 30 seconds, extend at 72°C for 1 minute), then the final extension at 72°C for 5 minutes before holding the plate at 4°C until the next stage of the protocol.

The reaction mix was analysed by gel electrophoresis to test for the correct sized PCR products. A 1% agarose gel is made up with BioLine agarose and TAE (Tris/Acetic acid/EDTA) with a 1:20000 dilution of ThermoFisher Scientific SYBR® Safe for UV visualisation of PCR products using a Bio-Rad Gel

Doc™ XR+ System and Quantity One® 1-D analysis software (version 4.6.9). A ThermoFisher Scientific GeneRuler 100 bp DNA Ladder (cat. SM024) was used as a reference.

2.7.3 BigDye and Sanger Sequencing

Excess primers and dNTPs were cleaned from PCR products using 1µl of ExoSap (made from Exonuclease I (New England Biolabs) and shrimp alkaline phosphatase (GE Healthcare) at a 1:2 ratio) and incubated at 37°C for 60 minutes. The ExoSap was then deactivated at 80°C for 15 minutes.

Bidirectional Sanger sequencing was run using 2µl of purified PCR product, 0.75µl of BigDye (v3.1), 2µl 5x BigDye sequencing buffer, 1µl of either forward or reverse primer (10µM), and 4.25µl of H_2O. The semi-skirted reaction plates were placed on a BioRad Tetrad 2 (MJ Research DNA Engine Tetrad PTC-225) and for 25 cycles of; denaturing at 96°C for 10 seconds, annealing at 50°C for 5 seconds and extend at 60°C for 3 minutes and 30 seconds.

The BigDye reactions were cleaned with 40µl of 75% isopropanol and incubated at room temperature for 30 minutes. The sealed plate was centrifuged at 2092RCF for 45 minutes. The open plate was then inverted on to absorbent paper to remove the supernatant and placed inverted into a centrifuge at 33RCF for 30 seconds. The open plate was air dried for 10 minutes in a dark box or drawer.

Each reaction pellet was resuspended in 10µl of Hi-Di™ Formamide (Applied Biosystems) and 10µl of H_2O prior to analysis on an ABI sequence analyser (Applied Biosystems).

3 Investigating predisposition to CDH1-negative hereditary diffuse gastric cancer

3.1 Introductory statement

This work is adapted from the following publication:

Fewings, E., Larionov, A., Redman, J., Goldgraben, M.A., Scarth, J., Richardson, S., Brewer, C., Davidson, R., Ellis, I., Evans, D.G., et al. (2018). Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. Lancet Gastroenterol. Hepatol. 3, 489–498.

Library preparation and sequencing of the germline DNA samples described in this chapter was performed by James Redman, the Cancer Research UK Cambridge Institute Genomics Core, and the Department of Medical Genetics Stratified Medicine Core Laboratory. Sequencing data were downloaded and processed by me using an in-house WES pipeline generated by Alexey Larionov. All data analysis was designed and performed by me. Variant validation by Sanger sequencing was performed by me.

3.2 Abstract

The development of hereditary diffuse gastric cancer (HDGC) syndrome is associated with germline pathogenic variants in the E-cadherin gene *CDH1*. The risk assessment and management of HDGC families that do not carry a *CDH1* variant is restricted. It is therefore difficult for such families to make informed choices about surveillance and risk reducing surgery. This study aimed to identify new candidate genes for HDGC predisposition in families with no detected pathogenic *CDH1* variants (*CDH1*-NPV). Whole exome sequencing was performed on DNA extracted from blood obtained as part of the Familial Gastric Cancer Study. Analysis was performed across 39 individuals (28 affected and 11 unaffected) from 22 *CDH1*-NPV families that fulfil the international criteria for HDGC. Genes with loss-of-function variants were prioritised using gene interaction analysis to identify clusters of genes that could be involved in HDGC predisposition. Germline variants were identified in known cancer predisposition genes or lesser studied DNA repair genes in six HDGC families. A frameshift deletion within *PALB2* was found in a family with a history of gastric and breast cancer. Two *MSH2* variants were identified, one frameshift insertion and one previously described start loss, in unrelated affected individuals. One family was identified with a unique combination of variants in DNA repair genes *ATR* and *NBN*. A missense variant and a splice acceptor variant were seen in two unrelated families in DNA repair gene *RECQL5*. This study expands the role of known cancer predisposition genes *PALB2* and *MSH2* in the HDGC syndrome. It also puts forward new candidates in relation to HDGC risk within *CDH1*-NPV families.

3.3 Introduction

3.3.1 Hereditary Diffuse Gastric Cancer

Gastric cancer is the fourth most common cancer globally and the best characterised autosomal dominant inherited gastric cancer is the diffuse type, the hallmark of which is the presence of multiple foci of signet-ring cells (Guilford *et al.*, 2007). The term hereditary diffuse gastric cancer (HDGC) is applied to families with a history of diffuse gastric cancer (DGC) that meet one of the following criteria:(Caldas *et al.*, 1999; van der Post *et al.*, 2015)

a) At least two cases of GC in first or second-degree relatives regardless of age of onset, with one confirmed case of DGC.

b) One case of DGC diagnosed before the age of 40.

c) Any family history of DGC and lobular BC diagnosed before the age of 50.

3.3.2 Genetic predisposition to HDGC

Germline pathogenic variants in the E-cadherin gene (*CDH1*) explain 25-30% of HDGC cases with over 100 pathogenic germline variants currently described within this gene (Hansford *et al.*, 2015). E-cadherin is a transmembrane protein that interacts with the actin cytoskeleton via α-catenin, β-catenin and γ-catenin to maintain cell adhesion and differentiation (Sereno *et al.*, 2011). The cumulative lifetime risk of gastric cancer in *CDH1* pathogenic variant carriers ranges from 40-60% in males and 63-83% in females (Hansford *et al.*, 2015). Female carriers are also at risk of developing lobular breast cancer, with breast cancer surveillance also recommended to *CDH1* pathogenic variant carrying families.

Families with a strong history of DGC or lobular breast cancer, usually in combination with early onset cases, are recommended for *CDH1* testing. Families that meet such testing criteria are often defined as HDGC disorder families regardless of whether a known pathogenic *CDH1* variant is identified. For HDGC families with known pathogenic *CDH1* variants, there are guidelines for risk assessment, disease management and surveillance. Risk reducing therapy such as prophylactic gastrectomy is often offered to pathogenic variant carriers, and completely eliminates risk of gastric cancer (van der Post *et al.*, 2015), however the effect that surgery has on the patients quality of life should be considered and only offered where a true risk of DGC development has been ascertained. Because of this, prophylactic gastrectomies are rarely recommended to families without *CDH1* pathogenic variants.

Families that meet *CDH1* testing criteria but are negative for pathogenic *CDH1* variants or decide not to undergo or to delay surgery are offered regular screening by endoscopy (Fitzgerald *et al.*, 2010; Mi *et al.*, 2017). Individuals are tested for the presence of microscopic signet ring cells, which are treated as a precursor lesion to DGC. Individuals in which these precursor legions are identified are referred for a gastrectomy, however the nature of such lesions means that it is possible that foci remain undetected by endoscopy and random biopsy (van der Post *et al.*, 2015). For *CDH1*-no pathogenic variant (NPV) families, risk assessment is uncertain and therefore the efficacy of risk reducing strategies

is harder to assess. These families are therefore reliant on endoscopies to decide their treatment options, putting them at higher risk of developing undetected DGC.

Other familial cancer syndromes that have been linked to gastric cancer predisposition include Lynch syndrome, which is categorised by pathogenic variants in DNA mismatch repair genes, Peutz-Jeghers syndrome caused by pathogenic variants in *STK11,* and Li-Fraumeni syndrome which is associated with germline pathogenic *TP53* variants (Masciari *et al.*, 2011; Sereno *et al.*, 2011; van Lier *et al.*, 2011; van der Post *et al.*, 2015). DGC does not appear to be overrepresented in these syndromes, although this has not been comprehensively studied.

Lynch syndrome is commonly associated with an increased risk of colorectal cancer and is therefore often known as hereditary nonpolyposis colorectal cancer (HNPCC). Mismatch repair genes associated with this syndrome are *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM.* Pathogenic variants in this pathway lead to an instability of microsatellite repeat regions which is often detected in tumours by microsatellite instability (MSI) testing. As well as conferring a risk of colorectal cancer, Lynch syndrome variant carriers are also at risk of developing gastric cancer, with *MLH1* and *MSH2* pathogenic variants carrying a 4.8% and 9% risk respectively (Oliveira *et al.*, 2015).

Additionally, predicted pathogenic variants in the DNA double strand break repair genes *ATM*, *BRCA2,* and *PALB2* have been identified in HDGC families (Hansford *et al.*, 2015; Sahasrabudhe *et al.*, 2017). However, given the rarity of the findings, the associated DGC risk is hard to quantify, and therefore these variants are not used in routine clinical testing to aid identification of HDGC families.

3.3.3 Environmental and lifestyle risk factors

Of the known environmental risk factors to gastric cancer, infection with *Helicobacter pylori* (*H.pylori*) is the most well studied and carries around a six-fold increase in gastric cancer risk (Vogelaar *et al.*, 2017). Within developed countries, gastric cancer incidence associated with *H.pylori* infection has been cut by improving sanitation and food refrigeration (Lee and Derakhshan, 2013). However estimated infection rates from the World Gastroenterology Organization database (Hunt *et al.*, 2011) remain high in Africa and South America, in particular in Brazil where it is predicted that 82% of the population carry an *H.pylori* infection and gastric cancer incidence is estimated to be higher than average at 10.9 (world age-standardized rates per 100,000) (Lee and Derakhshan, 2013).

Smoking and alcohol consumption have also been associated with an increased risk of many different cancer types. Studies across Europe suggest that 17.6% of gastric cancers were associated with tobacco smoking (González *et al.*, 2003). No difference exists between smoking and gastric cancer risk in Caucasian and Asian populations (LaTorre *et al.*, 2009). There is currently no definitive evidence to suggest that heavy alcohol consumption alone increases gastric cancer risk, as it is suggested that smoking and dietary habits are confounding factors in these analyses (Tramacere *et al.*, 2012).

Dietary factors including red and processed meat and fat consumption increase risk of developing gastric cancer. This is particularly the case for *H.pylori* infected individuals in the top quartile of meat consumption who have an absolute risk of gastric cancer development of 0.3% in ten years. Fatty acids including oleic acid, α-linolenic acid and di-homo-γ-linolenic acid were suggested by the EPIC-EURGAST study to also be associated with an increased risk of gastric cancer (Chajès, Jenab and Romieu, 2011).

3.3.4 Surveillance and survival

Endoscopic screening is the primary technique used to identify individuals that are at a risk of HDGC but with no identified pathogenic *CDH1* variant. Frequency of endoscopies depends on the predicted risk of the individual however annual screening is currently recommended to HDGC families to identify early signs of malignancies and prompt referral for a risk reducing gastrectomy (van der Post *et al.*, 2015). The procedure, using white-light examination with random and targeted biopsies of the mucosa, aims to detect microscopic signet ring cell carcinoma (SRCC) foci. Targeted biopsies look for pale areas of the mucosa and have been shown to have a 41.7% sensitivity for SRCC detection (Mi *et al.*, 2017). In comparison, random biopsies have a higher sensitivity for SRCC at around 75% (Mi *et al.*, 2017).

For patients with a known pathogenic *CDH1* variant, who decide to delay gastrectomy, regular endoscopies have been shown to identify SRCC in 61.1% of cases, with the majority of these (63.6%) being detected on the first endoscopy (Mi *et al.*, 2017). It has however been noted that upon gastrectomy, 100% of samples have multiple SRCC foci, including those that appeared normal after high-magnification endoscopy, and multiple gastric biopsies (Chun *et al.*, 2001; Norton *et al.*, 2007). Additionally, due to the penetrance of *CDH1* pathogenic variants in HDGC, it is likely that SRCC will have developed prior to endoscopy. This suggests that although endoscopies are currently the primary method to assess the risk of invasive HDGC for an individual, regular endoscopies are still not recommended over a risk reducing gastrectomy for *CDH1* pathogenic variant carriers.

3.3.5 Features of diffuse gastric cancer

The development of SRCC is a known precursor to HDGC and is observed in endoscopies to ascertain an individual's risk of developing invasive disease. SRCC, named as such because the large amount of mucin in these cells pushes the nucleus to the periphery, causing the cells to resemble signet rings, can be defined as in situ or pagetoid depending on whether they are identified in the basal membrane or below the epithelium of glands (Hansford *et al.*, 2015). Once DGC has advanced, it presents as a poorly differentiated diffuse carcinoma that often causes a stiffening of the gastric wall (*linitis plastica*) (Hansford *et al.*, 2015).

The genetic characteristics of sporadic DGC tumours has been well studied, with one study showing 42.2% of early onset DGC tumours carry somatic pathogenic *CDH1* variants (Cho *et al.*, 2017). Within this set, a number of somatic variants were also identified in *TP53, ARID1A, KRAS, PIK3CA, ERBB3, TGFBR1, FBXW7, RHOA,* and *MAP2K1* (Cho *et al.*, 2017). TCGA data was previously analysed to

identify molecular characteristics of sporadic gastric cancer samples, distinguishing the tumours into those that were positive for Epstein-Barr virus (9%: 26 out of 295 samples), had a high occurrence of MSI (22%: 64 out of 295 samples), were genomically stable (20%: 58 out of 295 samples), and those that exhibited chromosomal instability (50%: 147 out of 295 samples)(Bass *et al.*, 2014). Within a subset of hypermutated tumours from this analysis, the genes *TP53, ARID1A, KRAS,* and *PIK3CA* were also somatically mutated (Bass *et al.*, 2014). This suggests that these genes are important in driving the development of invasive gastric carcinomas.

The sporadic diffuse cases within this set lie primarily in the genomically stable subtype. This subtype was enriched for a known fusion between *CLDN18* and *ARHGAP6 or ARHGAP26* that have been shown to cause a loss of epithelial integrity, leading to stomach H^+ leakage (Bass *et al.*, 2014; Yao *et al.*, 2015). Interestingly, within this subtype, 30% of cases had either a *CLDN18-ARHGAP* gene fusion or a *RHOA* pathogenic variant, however the two were never seen simultaneously, suggesting two main, distinct mechanisms of cancer progression that produce one phenotype (Bass *et al.*, 2014).

3.3.6 Aims

This study describes the whole exome sequencing of HDGC families with no identified pathogenic *CDH1* variants to identify DGC predisposition genes. The aims are as follows:

1. To explore the sequencing data of affected and unaffected individuals in HDGC families and identify likely disease associated variants.
2. To identify new genes or functional gene groups associated with risk of HDGC which hope to provide more accurate risk estimates for *CDH1*-negative DGC families.

3.4 Materials and Methods:

3.4.1 Study Population

Twenty-eight individuals diagnosed with DGC and eleven unaffected relatives were recruited from 22 DGC families (table 3.1) that had tested negative for *CDH1* pathogenic germline variants as part of an ethically approved study (MREC 97/5/32). Families (including first and second degree relatives) were categorised as having DGC syndrome based on the current criteria (Caldas *et al.*, 1999; Fitzgerald *et al.*, 2010; van der Post *et al.*, 2015). No individuals within this study were known to be heavy drinkers or smokers. As part of routine clinical testing, immunohistochemistry (IHC) staining for mismatch repair proteins was performed on tumours.

3.4.2 Whole exome sequencing and variant filtering

Germline DNA was extracted from blood or saliva and prepared for PE125 WES using the Nextera Rapid Capture Exome enrichment kit (Illumina). Sequencing was performed on HiSeq-4000 machines. VCF files were generated using a standard pipeline following GATK best practice recommendations for whole exome data (see Chapter 2: Methods for further details). Optimised hard filters were applied, including a VQSR truth sensitivity of 99.5% for SNPs and 97% for INDELs, an average 10x depth (variant DP) per sample and a QUAL threshold of 200. The QUAL threshold corresponded to a Ti/Tv ratio of 2 as calculated by Samtools VCF-Stats. Multi-allelic variants were flagged and excluded for the purpose of this analysis. Only genotypes with quality (GQ) >20 and individual depth (genotype DP) in sample < 500 were retained for further analysis.

The dataset was filtered to select 3,973 different uncommon (AF <0.05 in 1000 genomes), protein-affecting variants (loss of function, predicted deleterious and damaging missense (as flagged by SIFT and PolyPhen respectively) and inframe indels) that were observed in the 28 affected HDGC samples (HDGC affected allele count > 0). These intuitive filters aimed to remove variants that were least likely to be affecting HDGC predisposition. All candidates were manually examined for allele frequency (AF) in healthy controls, allele count (AC) in affected and unaffected individuals within families and protein affect further downstream.

Variants were aggregated into 2,847 genes which were filtered to select those that contain at least one loss of function variant. The top 1% most variable genes were also removed; this was determined by the number of rare, protein-affecting variants each gene contains within the set. A set of 732 genes (1,228 different variants) were retained for analysis. Variant and gene filtering steps are summarised in figure 3.1.

Scripts generated for all analysis downstream of VCF generation can be found at the following link (https://github.com/elliefewings/Fewings_HDGC_exome_2018). VCF data can be downloaded from the following repository (https://doi.org/10.17863/CAM.17181).

Family	Family ID	Number of samples sequenced		Diagnosis of proband (age)	Other cancers diagnosed in first or second degree relatives (age)	Candidate gene
		Affected	Unaffected			
1	GPQ_045	2	0	DGC (41)*	DGC (44)*, GC (57)	
2	GPQ_047	2	4	DGC (27)*	PnC, OvC (22), DGC (24)*, DGC (28)	
3	GPQ_048	1	0	DGC (40)*	GC (28), DGC (48)	
4	GST_172	1	2	DGC (55)*	BC, LC, LxC, DGC (44), DGC (52)	*PALB2*
5	GST_230	1	0	DGC (36) and CRC (47)*	DGC (37), LC (54), CRC (57), BC (50), DGC (61), DGC (79), LC (83)	
6	GST_256	1	0	DGC (37)*	BC, GC (63), GC (64)	*RECQL5*
7	GST_257	2	0	DGC (36)*	CRC, BC (43) and DGC (55)*	
8	GST_275	1	0	DGC (47) and LBC (36)*	GC (44)	*MSH2* **
9	GST_296	1	0	DGC (44)*	DGC (28)	
10	GST_345	1	2	DGC (28)*	BC, GC (44), GC (47)	
11	GST_349	4	1	DGC (23)*	SR*, SR*, BC (40s), DGC (45)*, PC (60s), CRC (75)	*ATR*, *NBN*
12	GST_358	1	0	DGC (68)*	LC, GC (49), GC (50), GC (76)	*MSH2* **
13	GST_368	1	0	DGC (47)*	GC, GC (50s), GC (60s)	
14	GST_440	1	0	DGC (23)*	DGC (40s), DGC (46), ThyC (30)	
15	GST_441	1	0	DGC (53)*	GC (49), GC (67), GC (71)	
16	GST_444	1	0	DGC (37)*	GC, BC (54), BC (65), CRC (66)	
17	GST_446	1	0	DGC (45)*	DGC (42)	
18	GST_455	1	0	DGC (48)*	GC (44), GC (54)	
19	GST_459	1	1	DGC (35)*	LC, UtC (65)	
20	GST_460	1	0	DGC (55)*	GC (51), CRC (76)	
21	GST_463	1	1	DGC (28)*	GC (53), BC (76), GC (80)	*RECQL5*
22	GST_464	1	0	DGC (30)*	GC, DGC (67)	

Table 3.1: The 22 families sequenced within this study, including diagnoses of the 28 affected individuals sequenced across the 22 families. First or second degree relatives of each proband that were diagnosed with cancer are described and those that were sequenced within this study are marked with *. In total, 39 individuals were sequenced, including 28 affected and 11 unaffected relatives.

* indicates sequenced affected individuals. ** No microsatellite instability was detected in tumour. Cancers described include: breast cancer (BC), colorectal cancer (CRC) , diffuse gastric cancer (DGC), gastric cancer (GC), laryngeal cancer (LxC), lung cancer (LC), ovarian cancer (OvC), peritoneal cancer (PnC) , prostate cancer (PC), signet ring cells (SR) , thyroid cancer (ThyC), uterine cancer (UtC)

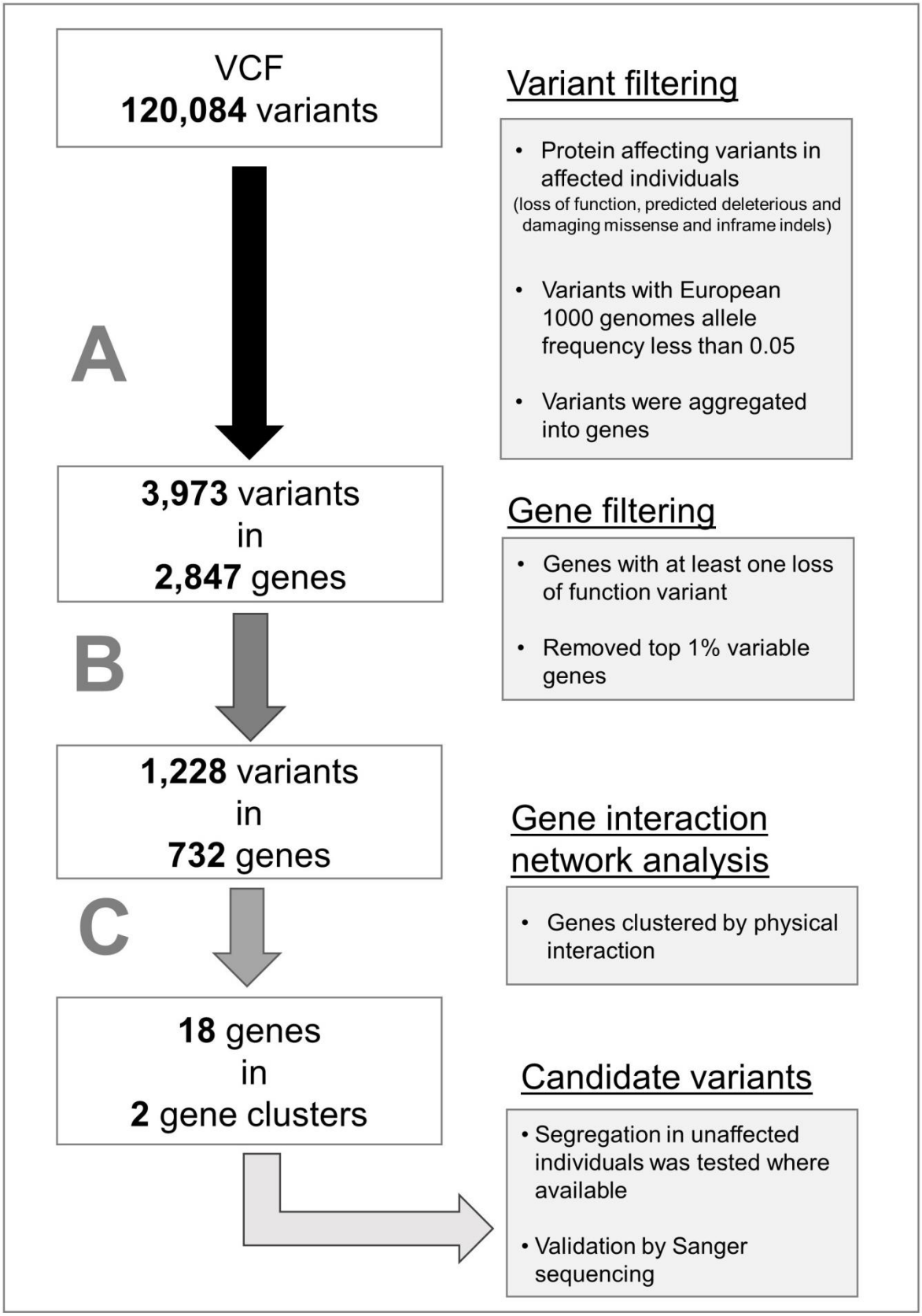


Figure 3.1: Variants filtering and analysis

A) variant filtering, B) gene filtering and C) gene clustering.

3.4.3 Gene interaction network analysis:

Gene interaction network analysis was used to identify variant-enriched candidate genes with interacting protein products; non-antagonistic, physically interacting proteins may have a similar effect on cell function and therefore may produce a shared phenotype when mutated. The 732 filtered genes were put through the Cytoscape GeneMania plugin, placing physically interacting genes into clusters for further analysis (Montojo *et al.*, 2010). A cluster was defined as a set of five or more physically interacting genes.

The Gene Ontology (GO) Consortium enrichment analysis web tool was used to apply GO terms to clusters using the PANTHER Overrepresentation Test (version 13.0) including the default Bonferroni correction for multiple testing (Blake *et al.*, 2015). Of the significant terms highlighted by the analysis, the most significant term that encompasses between ten and 200 genes was selected, in compliance with previous studies (Milne *et al.*, 2017).

Allelic counts of all 1,228 different filtered, loss of function variants (regardless of GeneMania clustering) within the selected GO terms were aggregated and contingency tables were drawn. Variants were also aggregated for each GO term over a comparably filtered set of 503 Europeans from phase-3 of the 1000 genomes study (Auton *et al.*, 2015). A one-tailed Fisher's exact test was performed using the R Stats package to test for an enrichment of loss of function variants within each selected GO term in HDGC in comparison to the European 1000 genomes set. For this test, only one occurrence of a variant was counted per family in the HDGC set.

The 1000 genomes project was used as a control set to test for an enrichment of loss of function variants under selected GO terms in HDGC. Variants from European phase-3 1000 genomes data were filtered to select 28,833 different uncommon (European AF <0.05 in 1000 genomes), protein-affecting variants (loss of function, predicted deleterious and damaging missense and inframe indels). Variants were aggregated into 11,796 genes, which were filtered to select those with at least one loss of function variant and remove the top 1% most variable genes. Variability was measured by the number of rare protein-affecting variants each gene contains; 3,634 genes containing 4,601 different loss of function variants were retained. Aggregated allele counts for each selected GO term were generated using these loss of function variants for further analysis.

3.4.4 Validation by Sanger sequencing

Candidate variants were validated by Sanger sequencing. Germline DNA from blood was quantified using the Qubit dsDNA HS kit (Invitrogen) and custom flanking primers were designed for each variant. DNA fragments were amplified by PCR and the products were sequenced on an ABI Genetic Analyser (Applied Biosystems) using BDT V3.1 (Invitrogen) according to the manufacturer's instructions for Sanger sequencing (see Chapter 2: Methods for further details). Due to their proximity, both RECQL5 variants (c.2806-2T>C and c.2828C>T) were covered by one pair of primers.

Tumour DNA from the father of family 11 was extracted from formalin-fixed paraffin embedded (FFPE) scrolls using a covaris ultrasonicator (Covaris). Identified germline candidates were checked for loss of heterozygosity within tumour DNA via Sanger sequencing as above.

3.4.5 Tumour immunohistochemistry and microsatellite instability analysis

The Ventana Benchmark mismatch repair panel (MLH-1 (M1), PMS2 (EPR3947), MSH2 (G219-1129) and CONFIRM anti-MSH6) was used to perform IHC analysis for known mismatch repair genes within selected tumours.

To perform MSI analysis, 5µm FFPE sections were mounted on glass slides for dewaxing and manual microdissection. DNA was extracted with the QIAamp DNA FFPE Tissue kit. Five standard microsatellite markers were used to evaluate the DNA (BAT-25, BAT-26, NR-21, NR-24, and MONO-27) using the Promega MSI Analysis System (v1.2). Poorly and moderately differentiated gastric tissue were compared to adjacent tumour-free tissue.

3.4.6 Analysis of *PALB2* and *BRCA2* variants in published studies

We searched PubMed without language restrictions between Jan 1, 2015, and Dec 31, 2017, using the term "hereditary diffuse gastric cancer" to identify sequencing studies reporting loss-of-function variants in *PALB2* and *BRCA2* in HDGC probands with no detected pathogenic *CDH1* variants. Only publications released after the initial report of *PALB2* and *BRCA2* in HDGC by Hansford et al were used (Hansford *et al.*, 2015). For each of the four identified publications (Hansford *et al.*, 2015; Sahasrabudhe *et al.*, 2017; Slavin *et al.*, 2017; Vogelaar *et al.*, 2017) and this current study, allele counts of loss of function *PALB2* and *BRCA2* variants were aggregated. The same counts were performed across the 503 European 1000 genomes samples and 27,173 non-TCGA non-Finnish Europeans from the ExAC control dataset (Lek *et al.*, 2016). Within all tested sets, the well-characterised *BRCA2* polymorphic stop codon in c.9976A>T was removed. A one-tailed Fisher's exact test was performed using the R Stats package to test for an enrichment of loss of function *PALB2* or *BRCA2* variants in HDGC in comparison to either control set.

3.4.7 Copy number variant analysis

All cases within this study were analysed for potential copy number variants (CNV). The XHMM algorithm, using principle component analysis (PCA) to normalise read depth over exomes and a hidden Markov model (HMM) to identify regions of increased or decreased reads that might indicate a CNV, was applied to the set of HDGC samples (Fromer *et al.*, 2012). All samples were analysed within their sequencing libraries to decrease the likelihood that CNVs were called due to variations in read depths generated during sequencing runs. Variable regions were annotated with the gene and exon that they covered for further analysis. Deletions were examined in BAM files to look for stretches of homozygosity within the called regions, as would appear if one allele had been lost. For duplications and deletions, the read depth of the called region was examined in comparison to surrounding regions and samples that were run in the same sequencing library. Around a 50% read drop or increase was

required for the variant to be considered for further analysis. CNVs were further explored in selected samples using an Affymetrix CytoScan 750K genotyping array according to clinical protocol.

3.4.8 Analysis of external data relating to candidate gene

Independently from the gene interaction analysis, variants that meet the above filters were manually explored to identify any additional candidates. One candidate variant was identified in *CLDN18*, which is implicated somatically in gastric cancer. Sequencing data from a study describing occurrences of the *CLDN18-ARHGAP26* fusion gene were downloaded using the SRA Toolkit, splitting files into two paired end FASTQs per sample (Leinonen, Sugawara and Shumway, 2011; Yao *et al.*, 2015). FASTQ files from three germline samples that were identified as carrying the somatic fusion gene were run through an in-house pipeline to generate aligned BAM files merged from multiple lanes of sequencing (see Chapter 2: Methods for further details). One sample failed analysis, BAM files from two others were explored manually in IGV to look for any variants in *CLDN18* that could be associated with a somatic *CLDN18-ARHGAP26* fusion gene.

3.5 Results

3.5.1 Gene interaction network analysis

A WES dataset of 39 individuals from 22 *CDH1*-NPV HDGC families (table 3.1) was filtered to select 3,973 different uncommon, protein-affecting variants that were aggregated into 2,847 genes. Genes were further filtered to select non-variable genes with at least one loss of function variant, creating a set of 732 genes (1,228 different variants). The presence of affected and unaffected family members within this set allowed for the selection of variants downstream that segregate with phenotype on a per family basis.

The set of 732 filtered genes was input into gene interaction network analysis, clustering 18 genes into two physical interaction clusters, shown in figure 3.2, to which GO terms were applied. A cluster of eight genes was identified (figure 3.2a) which was associated with the 'double strand break repair' (GO:0006302) GO term (PANTHER Overrepresentation Test p = 0.000000355). A second cluster of ten genes (figure 3.2b) was associated with the GO term 'negative regulation of extrinsic apoptotic signaling pathway via death domain receptors' (GO:1902042) (PANTHER Overrepresentation Test p value = 0.00517).

Loss of function variants within the filtered set of 1,228 different variants were aggregated under the two significant GO terms. This allowed the inclusion of related candidate genes that were not clustered by GeneMania. A one-tailed Fisher's exact test was applied to aggregated allelic counts for each GO term in the HDGC set and a comparably filtered European 1000 genomes set to look for an enrichment. The 'double strand break repair' term was significantly enriched in HDGC (p = 0.000512). In comparison, the 'apoptotic signaling pathway' term was not enriched in HDGC (p = 0.186), suggesting that the differences in allele counts seen in double strand break repair cannot entirely be explained by the technical differences that arise when using an externally produced control dataset.

The significantly enriched double strand break repair set was further explored to select candidate genes. *BRCA2* was a part of this set but was disregarded from further study as it contained the well-characterised, benign polymorphic stop codon c.9976A>T (Higgs *et al.*, 2015). Other genes in this set include *PALB2*, *MSH2*, *RECQL5*, *ATR*, and *NBN*, all of which were shown to be physically interacting in GeneMania. Candidate variants are summarised in table 3.2.

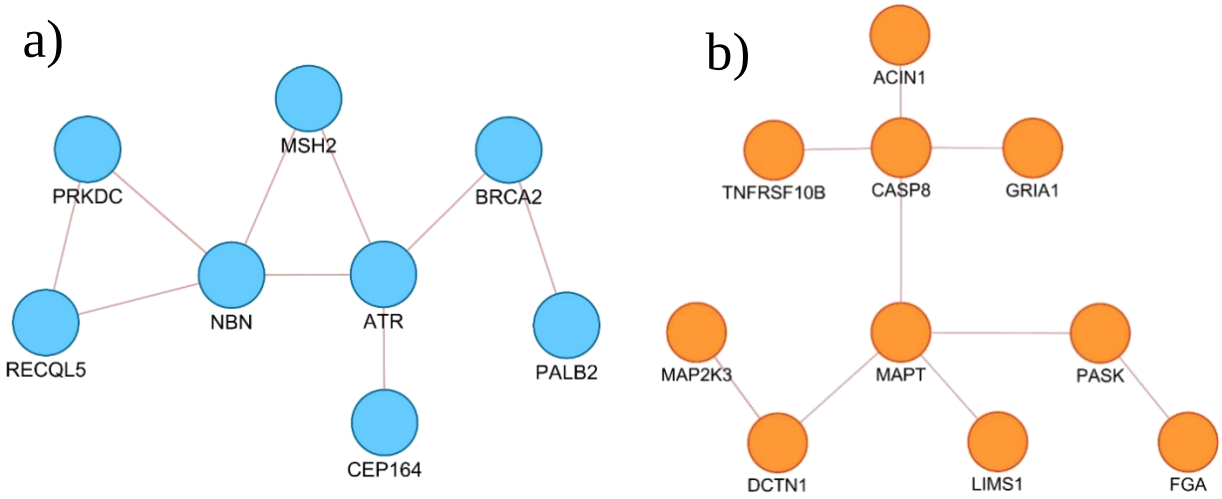


Figure 3.2: Gene clusters identified via gene interaction analysis

Pink lines indicate a physical interaction as assigned by the GeneMania plugin for Cytoscape (Montojo *et al.*, 2010). GO terms assigned to the clusters: a) 'double strand break repair' (GO:0006302), b) 'negative regulation of extrinsic apoptotic signaling pathway via death domain receptors' (GO:1902042).

Family	Number of sequenced individuals	Diagnosis of proband (age)	Other cancers diagnosed in first or second-degree relatives	Gene	Variant	Consequence	Protein Change	MAF 1000 genomes European	MAF ExAC non-TCGA European	SIFT	PolyPhen
4	3	DGC (55)	Breast, Lung, Larynx	***PALB2***	c.757_758delAG	Frameshift deletion	p.Leu253fs*	0	0	NA	NA
6	1	DGC (37)	Breast	***RECQL5***	c.2806-2T>C	Splice acceptor variant		0	0	NA	NA
8	1	DGC (47)	Breast	***MSH2***	c.967_968insCTCA	Frameshift insertion	p.Ser323fs*	0	0	NA	NA
11	5	DGC (28)	Prostate	***ATR***	c.6075A>T	Stop gain	p.Tyr2025*	0	0	NA	NA
				NBN	c.1124+1G>C	Splice donor variant		0	0	NA	NA
12	1	DGC (68)	Lung	***MSH2***	c.1A>C	Start loss	p.Met1?	0	0	Deleterious	Benign
21	2	DGC (28)	Breast	***RECQL5***	c.2828C>T	Missense variant	p.Arg943His	0.002	0.014332	Deleterious	Probably damaging

Table 3.2: Candidate variants identified via gene interaction analysis on WES data of HDGC families.

3.5.2 Candidate variants in HDGC families

A heterozygous 2bp frameshift deletion was identified in *PALB2* (NM_024675.3: c.757_758delAG, p.Leu253fs, rs180177092). This loss of function variant at amino acid position 253 is predicted to result in an early stop codon seven amino acids downstream. Family 4 (figure 3.3), in which this *PALB2* variant was found, has a history of breast, lung, and laryngeal cancer as well as DGC. Exome sequencing was performed on three siblings from this family; a proband with DGC at the age of 55 years, and two unaffected siblings. The variant was identified in the affected proband and in one of the siblings. The affected proband had previously received treatment for a *H. pylori* infection but had tested negative at all subsequent biopsies.

The mismatch repair gene *MSH2* is a member of the Lynch syndrome associated gene set. Within this gene, two loss of function variants were identified; a start loss (NM_000251.2: c.1A>C, p.Met1?, rs267607911) in family 12 and a frameshift insertion of 4bp (NM_000251.2: c.967_968insCTCA, p.Ser323fs) in family 8. Both variants were seen in families with a strong history of DGC, however only DNA from the probands was available for sequencing and segregation analysis could not be performed. Clinical IHC staining for mismatch repair proteins was performed on tumours from both cases, with both tumours showing normal expression levels of the MSH2 protein. Both families tested negative for *H. pylori*.

Variants in DNA repair genes *ATR* and *NBN* were seen within family 11 (figure 3.4). The proband of this family was diagnosed with DGC at age 28, two siblings underwent risk-reducing gastrectomies and were found to be signet ring cell positive. For the purpose of this analysis, these individuals are treated as affected family members. The family had an unusual history, with both parents being likely affected with DGC; the father was diagnosed with DGC and the mother had metastatic disease characterised by signet rings which were likely from a gastric primary. All individuals sequenced within this family tested negative for *H. pylori* infection. The proband and both siblings were double heterozygotes for loss of function variants in *ATR* and *NBN*. A splice donor variant (NM_002485.4: c.1124+1G>C) in *NBN* was seen in the three siblings but not the father and so was presumed to have been inherited maternally (DNA was only available for the father). A predicted stop gain variant (NM_001184.3: c.6075A>T, p.Tyr2025*) in *ATR* was seen in all three siblings and the father. Therefore, the three siblings carried the loss of function variants in both *NBN* and *ATR*. An unaffected second-degree relative within family 11 did not carry either variant. There were no other strong candidates within this family and a clinical Affymetrix CytoScan 750K SNP genotyping array to detect copy number changes in the father revealed no clinically relevant variants. DNA from tumour tissue from the father of family 11 was analysed for loss of heterozygosity of the *ATR* stop gain variant. The variant was confirmed as appearing within the tumour but remained in heterozygous state.

Gene	Variant	Consequence	Protein Change	Sample 1	Sample 2	Sample 3
PALB2	c.757-758TAG>T	Frameshift variant	p.Leu253fs	TAG/T	TAG/T	TAG/TAG

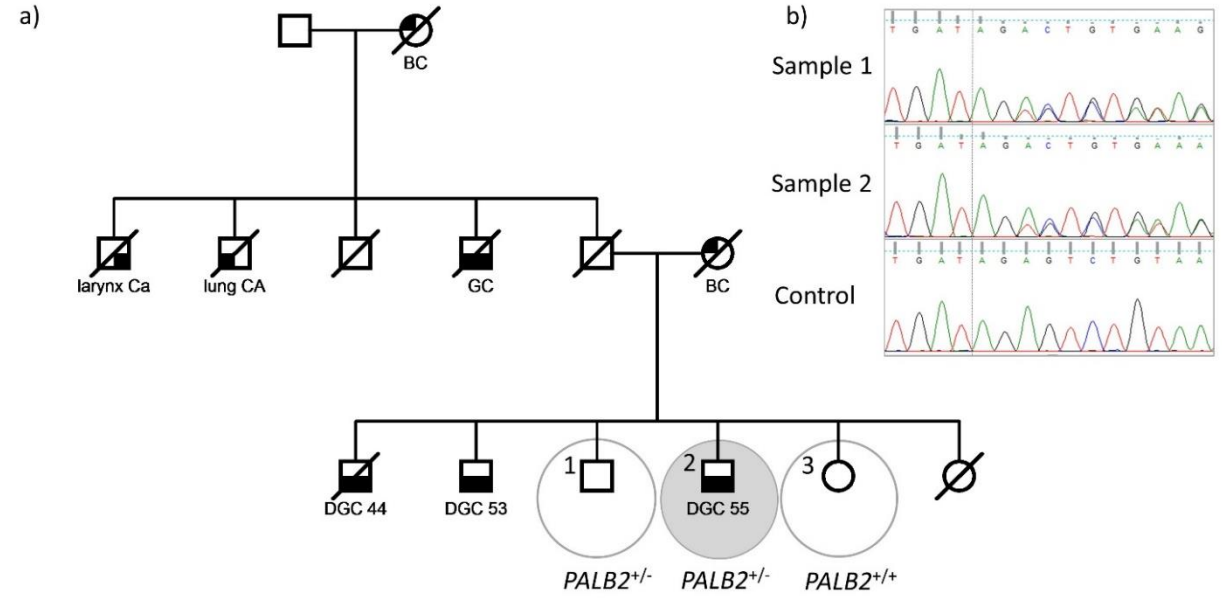


Figure 3.3: a) The pedigree for family 4. b) Chromatograms showing the frameshift variant in sample 1 and sample 2 against control DNA. Whole exome sequencing was performed on the circled samples, where shading indicates an affected and white indicates an unaffected individual.

Gene	Variant	Consequence	Protein Change	Sample 1	Sample 2	Sample 3	Sample 4
ATR	c.6075A>T	Stop gain	p.Tyr2025Ter	A/T	A/T	A/T	A/T
NBN	c.1124+1G>C	Splice donor variant		G/G	G/C	G/C	G/C

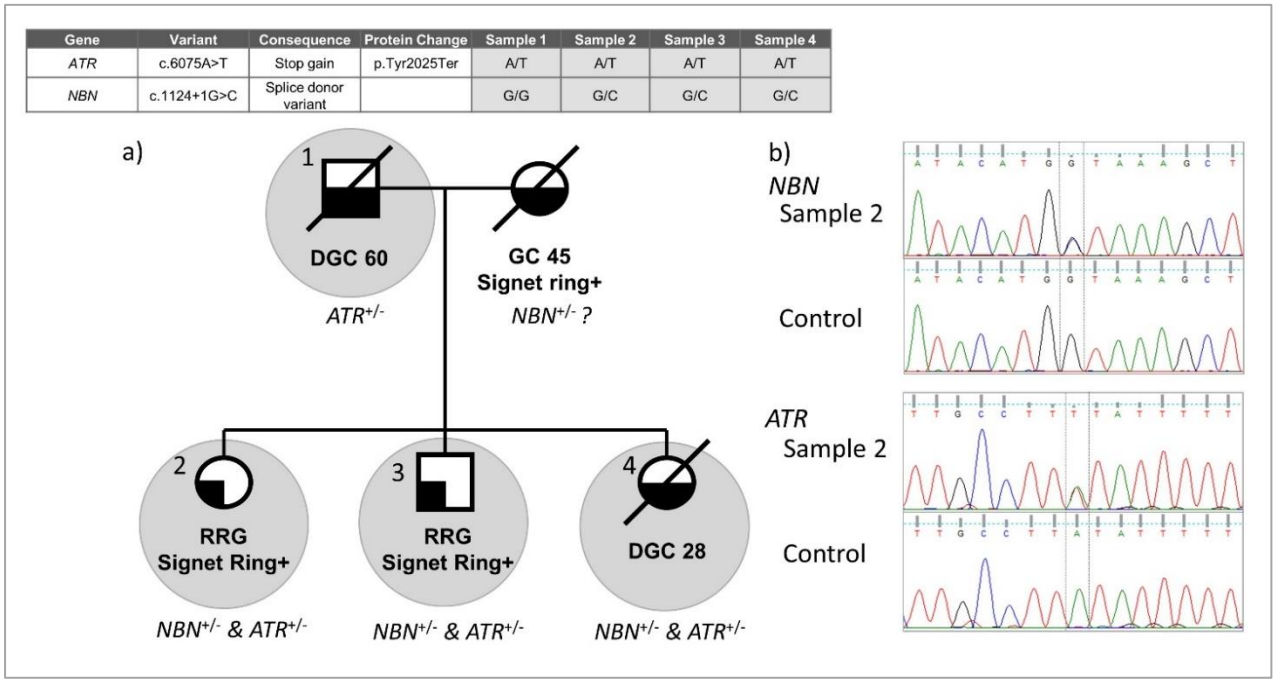


Figure 3.4: a) The pedigree for family 11 showing the segregation of variants in *NBN* and *ATR* amongst the four affected family members for which DNA was available for sequencing. b) Chromatograms showing the *NBN* and ATR variants in sample 2 against control DNA. Whole exome sequencing was performed on the circled samples, where shading indicates an affected individual.

Two variants were identified in the helicase gene *RECQL5* in different families; one missense variant (NM_004259.6: c.2828C>T, p.Arg943His, rs200535477) in family 21 and one loss of function splice acceptor variant (NM_004259.6: c.2806-2T>C, rs201841487) in family 6. In both families, gastric cancer and breast cancer can be seen across three generations. The proband of family 6 tested negative for *H. pylori*. The *H. pylori* status of the proband from family 21 is unknown. Within family 21 sequencing was performed on both an affected and unaffected individual, with the missense variant only appearing in the affected individual.

Other genes within the GeneMania cluster were not explored further as the genotype of individual variants did not segregate with the phenotype in families containing affected and unaffected members.

3.5.3 Loss of function variants in *PALB2* and *BRCA2* in published HDGC studies

The enrichment of loss of function *PALB2* and *BRCA2* variants in HDGC was assessed using previously published studies (Hansford *et al.*, 2015; Sahasrabudhe *et al.*, 2017; Slavin *et al.*, 2017; Vogelaar *et al.*, 2017). In total, five HDGC probands were identified with loss of function *PALB2* variants, out of 312 tested families across the four studies (table 3.3). In these studies, *PALB2* variant carriers account for 1.5% of HDGC families; in comparison to 0.096% of non-TCGA, non-Finnish European ExAC samples (Fisher's Exact Test, $p = 1.43x10^{-9}$) and 0.2% of European 1000 genomes samples ($p = 0.039$). In contrast, there was no enrichment of *BRCA2* loss of function variants in HDGC probands in comparison to ExAC ($p = 0.47$) and 1000 genomes ($p = 1$).

3.5.4 Germline copy number variants in HDGC

CNVs were called across the HDGC set and variants that lie in DNA repair or known cancer genes were further analysed. A duplication of exons 19 to 27 of DNA repair gene *ATM* was called in the one individual from family 9 and the affected individual from family 20 was called with a duplication of exons 1 to 15 of HDGC predisposition gene *CDH1*. A deletion of 16 exons within *FANCD2* was called in the one individual in family 13 however was later removed as a candidate due to the identification of heterozygous variants in called regions which are not consistent with a deletion.

The called duplications were explored in BAM files, neither showed a 50% gain in reads across the region in comparison to other germline DNA samples sequenced within the library. However, the link between *CDH1* and HDGC predisposition and its potential clinical implications warranted further validation of this variant within family 20 using a clinical grade SNP genotyping array. The *CDH1* duplication was not reproduced within this sample however a gain of approximately 83KB within 22q11.23 was identified.

The father from family 11 was also put forward for a SNP array to detect CNVs due to the likely inherited nature of disease within this family. A gain of approximately 1.5MB, spanning 2q12.2 and 2q12.3, and a hemizygous loss of 71Kb within Xp11.22 were detected. Neither covered regions were deemed clinically relevant to the phenotype.

Study	Race/Ethnicity	ID	Diagnosis of proband	Tumour work completed?	Variant	Consequence	Protein Change	MAF 1000 genomes European	MAF ExAC non-TCGA European
Hansford 2015	European	P124	DGC (45)	No	c.1193AC>A	Frameshift deletion	p.Val398fs	0	0
Sahasrabudhe 2017	European	CG-12	IGC (69)	Yes	c.1240C>T	Stop gain	p.Arg414Ter	0	0
Sahasrabudhe 2017	European	CG-008	DGC (48)	No	c.1240C>T	Stop gain	p.Arg414Ter	0	0
Sahasrabudhe 2017	European	GM037589	NA (46)	No	c.1240C>T	Stop gain	p.Arg414Ter	0	0
Sahasrabudhe 2017	European	CG-05	DGC (50)	No	c.3201+1G>T	Splice site variant		0	0
Sahasrabudhe 2017	European	CG-039	DGC (47)	No	c.1882_1890delAAGTCCTGC	Inframe deletion	p.Lys628_Cys630del	0	0
Sahasrabudhe 2017	Latin_american	CG-028	IGC (81)	Yes	c.1882_1890delAAGTCCTGC	Inframe deletion	p.Lys628_Cys630del	0	0
Sahasrabudhe 2017	Latin_american	3CG-103	Mixed (79)	Yes	c.2753C>A	Missense	p.Pro918Gln	0	0
Fewings 2018	European	GST_172_301	DGC (55)	No	c.757_758TAG>T	Frameshift deletion	p.Leu253fs	0	0
Teixeira (unpublished)	European	GM048157	DGC (56)	No	c.1438A>T	Stop gain	p.Lys480Ter	0	0

Table 3.3: *PALB2* variants identified by in HDGC sequencing studies

DGC= Diffuse Gastric Cancer, IGC = Intestinal Gastric Cancer

3.5.5 Analysing candidate variants in external data

Two affected individuals within family 7 were identified as carrying a germline start loss of gene *CLDN18* (NM_001002026.2: c.1A>G, p.Met1?, rs138856042). The gene was selected as a candidate due to the frequency of *CLDN18* fusion genes identified in gastric tumours (Bass *et al.*, 2014; Yao *et al.*, 2015). Three germline DNA samples from external data that were known to contain somatic *CLDN18* fusion genes were downloaded. One sample failed analysis, likely due to insufficient coverage, and the two other samples were explored for germline *CLDN18* variants. No variants were identified in *CLDN18* exonic regions of these two samples. It was therefore assumed that somatic fusion genes involving *CLDN18* were not a result of genomic instability in this gene caused by germline variants. This variant was therefore no longer considered a candidate for HDGC predisposition within this family.

3.6 Discussion

This study analysed the sequencing data of 28 individuals affected with HDGC. The number of affected individuals recruited to this study and therefore the statistical power to detect disease associated variants was greatly limited by the rarity of the disease. The inclusion of unaffected relatives made it possible to discern whether variants of interest segregated with the phenotype. However to add statistical evidence to the identified candidate variants, larger sample numbers would need to be sequenced to generate an extensive affected disease cohort. This could be achieved through the recruitment of additional affected individuals but this may not be achievable to any great number within a reasonable time frame. Alternatively, externally sequenced data can be combined into one analysis, allowing the benefit of larger sample sizes but introducing ethnic genetic differences that would need to be accounted for in association analyses.

Predicted pathogenic germline variants in known cancer predisposing genes or DNA repair genes, including PALB2, MSH2, ATR, NBN, and RECQL5, were found in 25% of HDGC families analysed in this study and these findings reflect the on-going expansion of cancer phenotypes in existing predisposition genes as genomic analyses extend to rarer cancer subtypes. Genes that were once seen as causing predisposition in one or two specific cancer types are now being further identified in the context of other cancers. This is largely the case for DNA damage repair genes, although specific repair pathways are closely associated to a few predominant cancer types. For example, mismatch repair genes were initially associated with colorectal cancer, but then subsequently associated with a risk of developing gastric and pancreatic cancer, amongst others (Lynch *et al.*, 1985; Oliveira *et al.*, 2015). However, simply identifying predicted pathogenic variants in known cancer predisposing genes does not imply causality. For this, larger studies with matched controls are required, which is not always feasible for rare diseases.

The generation of large exome/genome datasets, such as ExAC and 1000 genomes, can be used to strengthen possible associations as we have illustrated in the case of *PALB2*. When combining our data with those from previously published relevant studies, including those where no *PALB2* variants were found, we saw a significant overrepresentation of *PALB2* (but not *BRCA2*) pathogenic variants compared to ExAC and 1000 genomes controls. The discrepancies in p values seen in both controls is likely due to differences in sample size. In addition to these published families, another HDGC family with a strong family history of DGC with a *PALB2* stop gain variant (NM_024675.3: c.1438A>T, p.Lys480*) was reported to us by collaborators during the time course of this study. In a number of the cases described by Sahasrabudhe et al, tumour molecular profiling was also performed, showing that *PALB2* pathogenic variant carriers had mutational signatures indicative of homologous recombination defects (Sahasrabudhe *et al.*, 2017). The PALB2 protein plays a critical role in homologous recombination during double strand break DNA repair by recruiting BRCA2 and RAD51 to DNA breaks. Pathogenic variants in this gene are associated with an increased risk of breast and pancreatic

cancer (Antoniou *et al.*, 2014; Pauty *et al.*, 2014; Easton *et al.*, 2015). Even within *PALB2* families, cases of DGC are likely to be rare and could be masked by a larger number of sporadic gastric adenocarcinomas, which means that associations with cancer subtypes may be missed in epidemiological studies of *PALB2* families unless the pathology on all reported cancers is known. An analogous development has recently occurred with *BRCA1* where a rare serous subtype of endometrial cancer has shown to be overrepresented in *BRCA1* variant carriers, an example of a novel cancer association in a gene has been intensively studied for over 20 years (Saule *et al.*, 2018).

ATR and NBN are involved in initiating the response to double strand DNA breaks. The DNA repair signalling gene *NBN* produces one of three proteins that form the MRN complex, alongside MRE11 and RAD50. The MRN complex is involved in the activation of the ataxia proteins ATM and ATR, both of which play roles in the further recruitment of DNA damage repair proteins, cell cycle regulation, and apoptosis. Loss of function variants are seen in *NBN* and *ATR* independently of each other in the parents of family 11, both of which had DGC or suspected DGC. The variants are co-inherited in all three siblings within the family. Independently, these variants may be conferring a risk to DGC, although an extensive maternal and paternal family history show no other incidences of gastric cancer and only one instance of late-onset prostate cancer. Slavin et al also identified a stop gain variant in *ATR* in an individual with intestinal type adenocarcinoma and a strong family history of gastric cancer (Slavin *et al.*, 2017).

Nijmegen breakage syndrome (NBS) is a recessive condition that is associated with homozygous variants in *NBN.* One specific founder variant, c.657del5 is homozygous in more than 90% of patients with NBS, and in heterozygous form has been linked to stomach, colorectal and pancreatic cancer predisposition (Seemanová *et al.*, 2007). Other heterozygous variants in *NBN* have also been linked with an increased risk of lung, ovarian, and breast cancer (Rainville and Rana, 2014; Maria Kałużna *et al.*, 2015; Ramus *et al.*, 2015). A large case-control study of a Chinese population identified a haplotype of three SNPs within the 3' untranslated region of *NBN* that confers a protective effect on gastric cancer (P. Sun *et al.*, 2015), suggesting that there is increased likelihood that *NBN* plays a role in gastric cancer susceptibility.

The unusual cancer pattern seen in family 11 could be attributed to multi-locus inherited neoplasia alleles syndrome (MINAS), in which pathogenic variants are inherited in multiple cancer predisposing genes, leading to a distinct or severe phenotype (Whitworth, Skytte, Sunde, Derek H Lim, *et al.*, 2016). The close relationship between NBN and ATR in double strand break repair could indicate a potential combinatorial effect of these loss of function variants. An indication of this could be the young age of onset of DGC or signet ring cells seen in the three siblings. However, the combinatorial effect of double heterozygosity of variants in DNA repair genes has been studied extensively in *BRCA1* and *BRCA2*, and this appears to be no more deleterious to breast cancer predisposition than a single pathogenic

variant (Leegte, 2005). Nevertheless, double heterozygosity may have implications for genetic counselling that should be considered.

Within families such as this, that show a similar phenotype in both parents, and within all three offspring, careful consideration of the inheritance is required for counselling. In a heterogenous population, it is unlikely that both parents are carrying the same predisposition variant, and therefore predisposition is unlikely to be conferred by a homozygous variant within the three siblings. With the exome sequencing of one parent, compound heterozygous variants can be examined, with one variant inherited from each parent. Within this family, no compound heterozygosity was seen in genes that could be linked to cancer predisposition. The inheritance of two variants in interacting genes is a phenomenon that has been described in several cases, including combinations of BRCA2/NF1 in breast ductal carcinomas with multiple other neoplasms (Whitworth, Skytte, Sunde, Derek H Lim, *et al.*, 2016), and APC/MSH2 pathogenic variants in colorectal cancer (Uhrhammer and Bignon, 2008). As more information is made available about gene interactions and cancer predisposing pathways it is likely that more families such as this are identified, and knowledge of cancer predisposition gene interactions will be broadened.

Genetic variants in the mismatch DNA repair pathway are associated with Lynch syndrome. There were two variants identified in *MSH2*, a frameshift insertion and a start loss variant. A similar *MSH2* initiation codon variant (c.1A>G) has been previously described as having a mild effect on protein function and so should not be treated as a loss of function variant (Kets *et al.*, 2009). However, the function of MSH2, measured by the presence of MSI, has been shown to be reduced in cases and tumours with this variant (Kets *et al.*, 2009). However, both cases shown here have normal expression of MSH2 after IHC staining and no MSI within the tumours. While it is most likely that these cases represent phenocopies, with neither tumour appearing to be caused by a mismatch repair deficiency, a novel (non-mismatch repair mediated) mechanism of MSH2-related carcinogenesis in DGC cannot be ruled out.

The helicase RECQL5 is important for preventing aberrant homologous recombination and preventing the accumulation of double strand DNA breaks and preserving genome stability (Saponaro *et al.*, 2014). Two individuals were sequenced within family 21, with the missense *RECQL5* variant segregating between the affected proband and the unaffected father. The splice acceptor variant in family 6 was seen in an individual diagnosed with DGC at age 37. Both families have a history of breast and gastric cancer.

3.7 Summary

Previous studies have explored the role of known cancer predisposition genes in *CDH1*-NPV HDGC. Sahasrabudhe et al identify *PALB2*, *BRCA1*, and *RAD51C* germline variants in DGC families (Sahasrabudhe *et al.*, 2017). Hansford et al describe variants in *ATM*, *BRCA2*, *MSR1*, and *STK11*, a frameshift deletion in *PALB2*, as well as uncovering a role for *CDH1* related adhesion gene *CTNNA1*

(Hansford *et al.*, 2015). Although our study found no variants of interest in *ATM, BRCA1, BRCA2, CTNNA1, MSR1, RAD51C,* or *STK11,* an exploration of *PALB2* variants in all *HDGC* families sequenced within recent literature showed an enrichment of loss of function variants in comparison to control sets. The results of these studies make a case for including *PALB2* in genetic testing for *CDH1*-NPV HDGC families and it is possible that *PALB2* pathogenic variant carriers with DGC may benefit from platinum-based chemotherapy and treatment with PARP inhibitors. On the other hand, the evidence is not yet strong enough to recommend DGC surveillance in *PALB2* pathogenic variant carriers, as the absolute level of risk is likely to be low in the absence of a family history.

Sporadic cancers, including a stomach adenocarcinoma subset had been analysed as part of a TCGA study, looking for rare germline variants in various cancer types (Lu *et al.*, 2015). This study identified an association between truncating *PALB2* variants and sporadic stomach adenocarcinoma (Lu *et al.*, 2015). To supplement this study, TCGA data were analysed, selecting 88 DGC individuals as described by Bass et al (Bass *et al.*, 2014). No truncating *PALB2* variants were seen within sporadic DGC samples, suggesting that the described association with *PALB2* truncating variants is seen within intestinal gastric cancer samples and so may also play a role in sporadic intestinal gastric cancers.

The rarity of *CDH1*-NPV HDGC makes the collection of large datasets challenging. This study employed gene interaction network analysis to prioritise candidate variants that are most likely to be involved in HDGC predisposition based on prior biological knowledge. This does not overcome the problem of low statistical power due to the small sample size that is often seen with rare cancer datasets. However, this method allows for the selection of the most plausible candidates from the available data.

An additional analysis of CNVs within this dataset was attempted using the XHMM algorithm (Fromer *et al.*, 2012). Although it has not proposed any plausible candidates, the CNV analysis on germline WES data is currently not validated and therefore some causal CNVs could have been missed within this study.

In summary, DNA damage repair genes were found to be enriched for rare protein-affecting variants within HDGC families with no detected pathogenic *CDH1* variants (*CDH1*-NPV). Further studies of these genes in similar families are required to gather more evidence on these potential associations, increasing knowledge about the genetic basis of these families to make better informed decisions about risk reduction and possibly influence management in affected family members. Lastly, many *CDH1*-NPV families remain unexplained even after WES, and while WGS may identify some further candidates in regulatory elements or structural variants it seems unlikely that a second high-impact gene implicated in HDGC will be discovered beyond *CDH1*. With this in mind, focusing on moderate or low impact cancer genes such as *PALB2* may be the way forward for studies into *CDH1*-NPV HDGC predisposition.

4 MALTA (*MYH9* Associated eLasTin Aggregation) syndrome: germline variants in *MYH9* cause rare sweat duct proliferations and irregular elastin aggregations

4.1 Introductory statement

The work described in this chapter is the culmination of clinical and bioinformatics research. The separation of this phenotype from previously described sweat duct proliferative phenotypes was noted by Schaller et al (Schaller *et al.*, 2010). Cases 1, 2-1, and 2-2 were described by Jörg Schaller and Ed Rytina. WES on these samples was performed by the Cancer Research UK Cambridge Institute Genomics Core. Data were processed to VCF by an in-house WES pipeline (written by Alexey Larionov) and analysed by me. Family 3 was identified, sequenced and analysed by Mirjana Ziemer, Konstanze Hörtnagel, and Kerstin Reicherter who highlighted a number of candidate genes. Data were then shared with me for joint variant calling and analysis. The candidate gene was sequenced in families 4 and 5 by Mirjana Ziemer, Konstanze Hörtnagel, and Kerstin Reicherter. All *in silico* functional work on the identified variants was designed and performed by me.

4.2 Abstract

In the present study we investigate the germline whole exome sequencing data of several individuals with an irregular distribution of elastin fibres and benign sweat duct proliferations who were initially diagnosed with microcystic adnexal carcinoma, Rombo or Nicolau-Balus syndrome. Germline DNA was extracted from the blood of six affected individuals from three families with a history of sweat duct proliferations. Whole exome sequencing was performed across these individuals to select candidate rare protein-affecting variants that co-segregate with phenotype. The candidate gene *MYH9* was Sanger sequenced in a further two affected families. Analysis of the germline variants within these individuals provides evidence that these disorders represent a single entity, characterised by variants in non-muscle myosin gene *MYH9*. Non-muscle myosins take part in a diverse range of cellular functions including cell migration and invasion, protein and organelle localisation and cell signalling. Pathogenic variants in *MYH9* have been previously described in patients with thrombocytopaenia and giant platelets. None of the affected individuals in this study presented with any phenotypic features of the previously described *MYH9*-related platelet disorder (MRPD). Four out of the five identified variants were located within the myosin head ATP binding pocket. The amino acids modified by these variants are conserved across eight different classes of myosin. We provide evidence that sweat duct proliferations with abnormal elastin aggregations are associated with germline pathogenic variants in *MYH9* and therefore suggest the name MALTA (MYH9 Associated eLasTin Aggregation) syndrome to reflect this underlying genetic basis. The identified *MYH9* variants are located in different gene regions and appear to have a different effect on myosin function to those previously associated with MRPD. Understanding the genetic basis of MALTA syndrome will aid the diagnosis and management of this disease by providing a way of differentiating its associated skin lesions from sweat duct tumours that have very

similar histopathological characteristics but are clinically aggressive. Importantly, our findings point to a novel disease mechanism for non-muscle myosins in tumorigenesis and skin diseases.

4.3 Introduction

In 1961 Nicolau and Balus described individuals with benign cutaneous lesions including atrophodermia vermiculata, multiple syringomata and milia which become known as Nicolau-Balus syndrome and, amongst other histopathological features, was characterised by irregular distribution of elastin fibres (Nicolau and Balus, 1961). A similar phenotype was later described as Rombo syndrome by Michaelsson et al (Michaelsson, Olsson and Westermark, 1981), with elastin distribution in these cases appearing like "swathes of steel wool" in some areas of the dermis (Van Steensel, Jaspers and Steijlen, 2001). In 2010, Schaller et al described a number of cases with the same abnormal elastin tissue, with the additional feature of sweat duct proliferations that are morphologically similar to those found in microcystic adnexal carcinoma (MAC) patients (Schaller *et al.*, 2010). MAC is a locally aggressive disease composing of sweat duct proliferations that invade perineurally and require surgical resection in many cases (Goldstein, Barr and Santa Cruz, 1982). In contrast, the condition described by Schaller et al followed a benign course with no perineural invasion (Schaller *et al.*, 2010) and the two described cases showed no disease progression 19 years after partial excision of the lesions (Schaller *et al.*, 2010). This suggests that while the ductal proliferations seen in these cases mimic MAC, the disease progression and histopathological features are more closely related to Rombo syndrome. A summary of these phenotypically related syndromes is given in figure 4.1.

Sweat duct proliferations and related cutaneous tumours pose a challenge for diagnosis and treatment as they have multiple overlapping histopathological features. Understanding the genetic landscape of these complex skin lesions would aid diagnosis, particularly in distinguishing them from more aggressive neoplasms, thereby altering the treatment and management. We provide evidence that sweat duct proliferations with abnormal elastin aggregations are associated with germline pathogenic variants in *MYH9* and suggest the name MALTA (MYH9 Associated eLasTin Aggregation) syndrome to reflect this.

4.3.1 Microcystic adnexal carcinoma associated cutaneous neoplasia

MAC is a cutaneous neoplasm that was first described in 1982 by Goldstein et al (Goldstein, Barr and Santa Cruz, 1982). Goldstein described the disease as a distinct clinicopathologic entity and detailed the histological differences between MAC and other known adnexal tumours (Goldstein, Barr and Santa Cruz, 1982). In contrast to other described adnexal tumours, MACs exhibit highly local aggressive behaviour, as well as being largely invasive into skeletal muscle and perineural spaces (Goldstein, Barr and Santa Cruz, 1982). This perineural involvement and the propensity for the lesions to appear on the upper lip distinguish MAC from other similar adnexal tumours (Goldstein, Barr and Santa Cruz, 1982). However, differential diagnosis of sweat gland tumours has remained problematic due to the lack of consensus in classification which is often due to inadequate biopsy size (Smith *et al.*, 2001; Cardoso and Calonje, 2015). In particular, desmoplastic trichoepithelioma (dTE), sclerosing and morphea-type basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and syringoma are frequently and

incorrectly diagnosed (Goldstein, Barr and Santa Cruz, 1982; Crowson, Magro and Mihm, 2006). However, MACs display ductal structures which differentiates them from dTE and are also unique in their invasion both perineurally and subcutaneously (Goldstein, Barr and Santa Cruz, 1982).

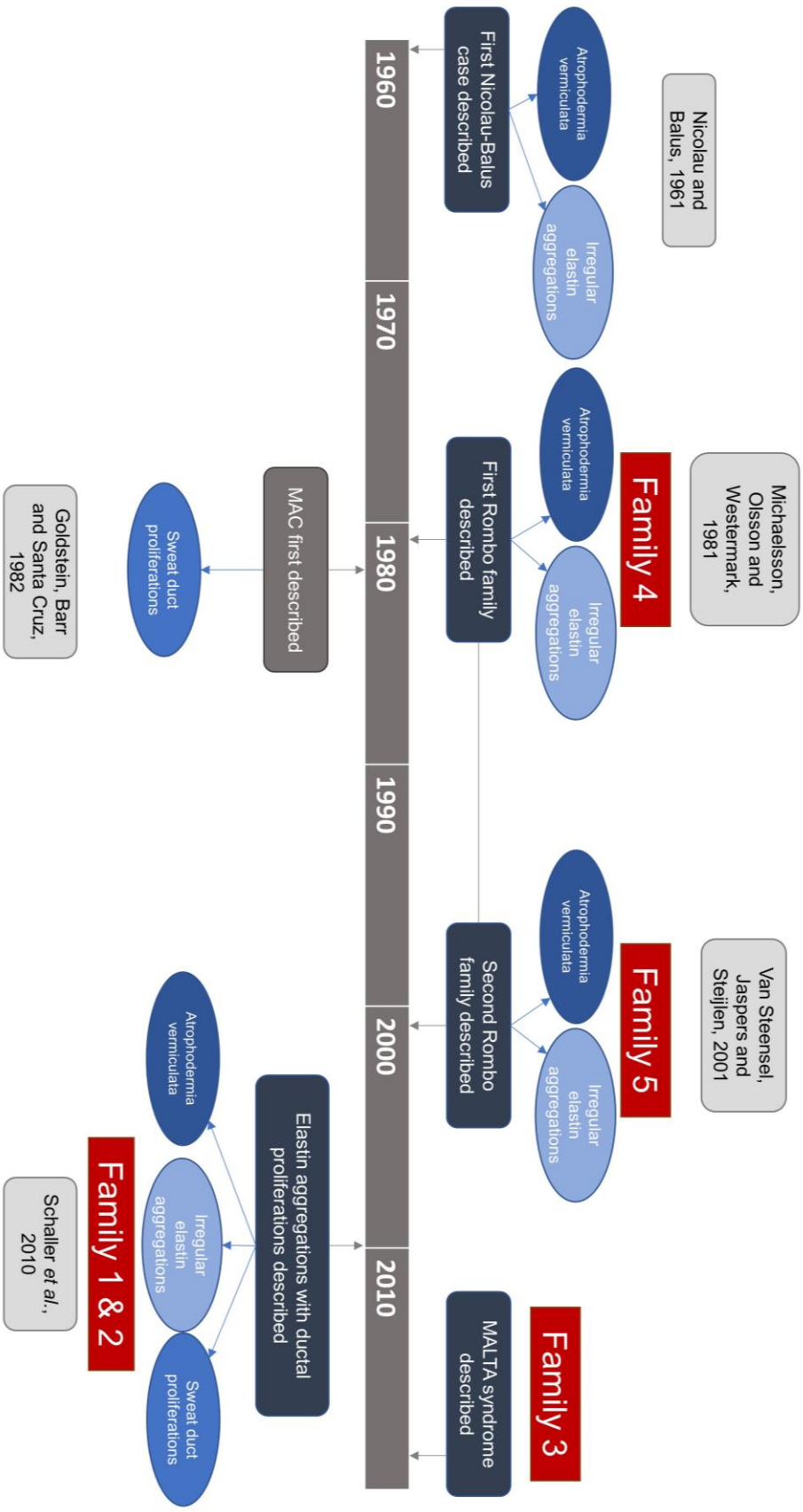


Figure 4.1: A timeline showing the identification of cutaneous neoplasms displaying atrophodermia vermiculata and elastin aggregations.

In spite of this, the literature suggests misdiagnosis is still a concern with related diseases, with one case study reporting that seven MAC cases were not diagnosed with MAC after an initial clinical examination (Gabillot-Carre *et al.*, 2006). Of these, six cases were misdiagnosed after further pathological examinations (Gabillot-Carre *et al.*, 2006).

MAC-like ductal structures have also been observed without such perineural invasion, and in combination with other features such as syringomata, atrophodermia vermiculata, milia and keratotic cysts (Schaller *et al.*, 2010). One main histopathological characteristic of these cases is an unusual aggregation and distribution of elastin fibres which have been described as ball like or wiry in various cases (Van Steensel, Jaspers and Steijlen, 2001; Schaller *et al.*, 2010). Such features are also described in Rombo and Nicolau-Balus syndrome however these cases have not been described with MAC-like ductal structures (Michaelsson, Olsson and Westermark, 1981; Van Steensel, Jaspers and Steijlen, 2001; Schaller *et al.*, 2010). Although these described cases have some overlapping features, such as elastin aggregations, other features are markedly less common. For example, the family described by Michaelsson et al were also identified with a number of BCC cases, developing at around the age of 35 (Michaelsson, Olsson and Westermark, 1981). Whilst other cases appear to be affected mainly superficially, causing social impairment in otherwise healthy individuals (Van Steensel, Jaspers and Steijlen, 2001). The spectrum of identified features in individuals with elastin aggregations can cause differential diagnosis. Within cases described by Schaller et al. a mother and daughter presented with different combinations of the above features and so were diagnosed with different diseases (Schaller *et al.*, 2010). In cases such as these, it is likely that the disease is inherited in an autosomal dominant manner and so, if possible, genotypic characterisation would more informative for diagnosis and management.

Case studies of Nicolau-Balus and Rombo syndrome described abnormal elastin distribution as a key diagnostic factor (Michaelsson, Olsson and Westermark, 1981). In these cases, in contrast to the families described within this study, no sweat duct proliferations were described (Michaelsson, Olsson and Westermark, 1981; Van Steensel, Jaspers and Steijlen, 2001; Schaller *et al.*, 2010). However it is possible that these were missed in the original biopsies, as was the case with one family described within this study within which sweat duct proliferations were identified upon retrospective review (Van Steensel, Jaspers and Steijlen, 2001). In contrast to this, Chiller et al (Chiller *et al.*, 2000) collated data from 48 MAC cases which were identified as having small keratinizing cysts and ductal proliferations but with no mention of elastin distribution. Therefore, although MALTA syndrome shows great similarities to MAC, and in some cases could be diagnosed as such, the irregular elastin distribution, relative benign nature of disease progression, and lack of MAC-like ductal proliferations in otherwise affected family members suggests this is a distinct entity.

Within this study we describe a benign syndrome that differentiates itself in its characteristic aggregations of elastin fibres and underlying genetic factor. Although the MAC-like ductal structures are a distinct cutaneous feature of this syndrome, they are not entirely descriptive of the disease as a whole and are not present in all cases. The common features that appear within described families include the keratinizing cysts and irregular distribution of elastin as is described in a mother and daughter by Schaller et al (Schaller *et al.*, 2010). The MAC-like ductal proliferations identified by Schaller et al (Schaller *et al.*, 2010) were only present in four of the five tested families.

4.3.2 Occurrences of adnexal neoplasms – Age, environment, gender and family history

The differential diagnoses of adnexal neoplasias limits the ability to generate a large scale epidemiological study of age and environmental factors. Such studies have been done for more specific phenotypes such as MAC, however these rely heavily on case studies to provide information about the clinical attributes of patients. Using case study clinical and histopathological information to study syndromes poses the challenge of making sure that all samples were correctly diagnosed and share the same phenotypic features. The MAC-like syndrome described within this study may be encompassed within some of the MAC cohorts, as at the time of diagnosis irregular distribution of elastin may have been missed by pathologist or noted as a sample irregularity during MAC diagnosis. Therefore, understanding the epidemiology of this and other similar adnexal neoplasms may give an indication of the occurrences of the described MAC-like syndrome.

A large scale epidemiological study of MAC occurrences has yet to take place. As such, individual case studies are heavily relied upon to provide information about the clinical attributes of patients. Some studies have collated individual sweat duct proliferation case studies to try and define groups at greater risk and to discuss the efficiency of treatments among these groups (Chiller *et al.*, 2000; Gabillot-Carre *et al.*, 2006; Schaller *et al.*, 2010). One benefit of such collation is that individual diagnoses can be discussed with reference to histopathological information from all participants, limiting differential diagnosis. An example of this is shown by Schaller et al. (Schaller *et al.*, 2010) in a table of patients from various studies shown in combination with their clinical and histopathological features. Another benefit is that factors such as gender, age of diagnosis, familial history and environmental exposures can be analysed to provide information which may aid in the diagnosis and management of the disease.

The average age of MAC patients is extremely diverse across studies. One large scale American study of 48 cases states the median age of diagnosis to be between 60 to 69 years (Chiller *et al.*, 2000). However they do note that their study includes one Male of 19-years-old and suggest a trend towards a younger diagnosis in male patients (Chiller *et al.*, 2000). Another report of seven cases similarly state their patients to be middle aged (Gabillot-Carre *et al.*, 2006). Interestingly, a summary of MAC-like cases by Schaller et al. gives an age of disease onset between 2 to 15 years of age (Schaller *et al.*, 2010). Another case study describing MAC on the axilla of an 18-year-old discusses this age difference between studies, suggesting that some lesions have been initially reported many years before a diagnosis

was ascertained (Green *et al.*, 2014). They go on to venture that the asymptomatic and indolent nature of MAC may lead to it being overlooked in children and young adolescents; and perhaps not be reported or diagnosed until later in life (Green *et al.*, 2014). This is emphasised by the presence of paediatric cases of MAC, with some being congenital (Green *et al.*, 2014), which suggests that genetic factors may be more influential than age and environmental exposure.

As with many cutaneous carcinomas, UV exposure is speculated to be a driving factor of MAC development. This speculation is coupled with the fact that MAC, as well as Rombo and similar syndromes, is often identified on the head or neck, and less frequently on the axilla (Crowson, Magro and Mihm, 2006). When this is considered along with a median diagnosis age of middle to late in life, it seems reasonable to suggest that the onset of the disease is coupled with an environmental factor or linked to aging. One American epidemiological study has suggested that a propensity for the lesion to occur on the left side of the face may be linked to increased UV exposure whilst driving (Chiller *et al.*, 2000). However there are noted occurrences of sweat duct proliferative lesions in regions that are not commonly UV exposed, including on the breast (Cardoso and Calonje, 2015) and buttocks (Schaller *et al.*, 2010). There are also several childhood and adolescent case studies (Schaller *et al.*, 2010; Green *et al.*, 2014), which suggests that genetic factors may contribute to risk in some cases.

In 2010 it was reported that fewer than 300 MAC cases had been described worldwide (Inskip and Magee, 2015). The rarity of this disease means that a wide scale study of the effects of UV exposure and genetic factors on MAC predisposition and development is currently not possible; although the effects of UV are often alluded to in the literature, with some case studies labelling exposure as a known predisposing factor (Inskip and Magee, 2015). When studying the effect of UV and genetic factors on MAC development it is important to make distinctions between middle to late life cases with no family history of cutaneous neoplasms, which are more likely to be sporadic and linked to environmental factors, and early onset cases or those with a family history of cutaneous neoplasms within which genetic factors are likely to play a role.

In contrast to many of the MAC cases, cases of Rombo syndrome were diagnosed during late childhood or adolescent years (Michaelsson, Olsson and Westermark, 1981; Van Steensel, Jaspers and Steijlen, 2001). Of the limited number of cases currently described, one is part of a family with a strong history of Rombo, with the syndrome being observed through four generations (Michaelsson, Olsson and Westermark, 1981). Cases within this family were first observed with skin conditions including atrophodermia vermiculata and milia-like papules that developed in late childhood (Michaelsson, Olsson and Westermark, 1981). A spontaneous case was presented by Van Steensel et al., this individual first presented with irregular facial skin at 15 years of age however the patient claimed to have begun noticing changes to the skin at around 6 years of age (Van Steensel, Jaspers and Steijlen, 2001). Ashinoff et al also identified a spontaneous Rombo patient who presented with atrophodermia

vermiculata, plaques on the left and right cheeks and no eyelashes, as was seen in previous cases (Michaelsson, Olsson and Westermark, 1981; Ashinoff, Jacobson and Belsito, 1993). Unusually, this patient was diagnosed at the age of around 94, and had apparently normal skin at the age of 35 (Ashinoff, Jacobson and Belsito, 1993). It has been further speculated that the case described by Ashinoff et al was not affected with Rombo syndrome due to a lack of described atrophodermia vermiculata as well as different histological features to those described by Michaelsson et al and Van Steensel et al (Michaelsson, Olsson and Westermark, 1981; Van Steensel, Jaspers and Steijlen, 2001).

For this study, which aims to identify an underlying genetic cause to MAC-like sweat duct proliferations in combination with irregular elastin aggregations, all samples were identified with an early age of onset or strong family history. This increases the likelihood that the identified phenotype is caused by underlying genetic causes. The development of similar disease characteristics later in life and with no family history may be indicative of spontaneous genetic variants or that these phenotypes can be mimicked by environmental predisposing factors.

4.3.3 Histopathology of MAC and MAC-like adnexal neoplasms

A report by Public Health England stated that there were 109 cases of MAC within the region from 2010 to 2013 (Public Health England, 2015). The rarity of such cancers means that each case study provides a large contribution to the field and a new insight into the disease. Consequently, the pathology of MAC is not singularly defined, but rather a collection of observations from various case studies. However, Goldstein et al (Goldstein, Barr and Santa Cruz, 1982) identified some core features of MAC that can distinguish this entity from its differential diagnoses, including the appearance of lesions on the upper lip and both perineural and subcutaneous involvement (Goldstein, Barr and Santa Cruz, 1982). MAC-like adnexal neoplasms may present with a number of the following features and may subsequently have been diagnosed as a different disease or syndrome.

MAC often presents as a plaque or subcutaneous nodule, which is usually within, but not limited to the head or neck region (Gabillot-Carre *et al.*, 2006; Cardoso and Calonje, 2015). Patient biopsy usually reveals keratocysts which present superficially and in the mid-dermis, alongside infiltrating ductal structures (Schaller *et al.*, 2010). These structures are surrounded by 2 layers of cuboidal cells in the deep dermis (McKinley *et al.*, 2014). Both MAC and MAC-like ductal proliferations are often also surrounded by myxoid connective tissue or collagenous stroma, giving an appearance that has been described as ‘onion-like’ (Schaller *et al.*, 2010; McKinley *et al.*, 2014).

One additional feature that has been described in the cases studied here is the aggregation of elastic tissue in the dermal papillae (Schaller *et al.*, 2010). Specifically, a number of cases show ball-like elastin structures within the papillary dermis (Schaller *et al.*, 2010). These aggregations are a feature of Nicolau-Balus syndrome and are often accompanied by reduced distribution of elastic fibres in the deep dermis which was shown in cases by Schaller et al. (Schaller *et al.*, 2010) with the addition of MAC-

like ductal proliferations. Interestingly, such distinct histopathological features which have been described in Rombo and Nicolau-Balus cases, have not been noted in previous MAC studies, where diagnosis is focused on the appearance of more classical ductal proliferations (LeBoit and Sexton, 1993; Chiller *et al.*, 2000; Cardoso and Calonje, 2015).

Additionally, MAC and MAC-like cases have been identified with syringomas (Schaller *et al.*, 2010; McKinley *et al.*, 2014; Cardoso and Calonje, 2015). These benign neoplasms of eccrine origin have been described as yellow or flesh coloured papules and can appear as an individual entity or in conjunction with another disease (Lau and Haber, 2013). It has been proposed that some patients with MAC-like ductal proliferations may instead be suffering from a rare form of 'plaque-like' syringoma, such is the similarity of the histopathology (Schaller *et al.*, 2010). This is supported by the presence of cuboidal epithelial cells with sclerotic collagen (Schaller *et al.*, 2010) as seen in cases of eruptive syringoma (Cannon, 1981; Lau and Haber, 2013).

Cases of Rombo syndrome have been described with similar histopathological features. One case study reports cysts in the middle dermis containing vellus hairs and horny material, with a general thickening of the stratum corneum identified from biopsies and described as hyperorthokeratosis (Van Steensel, Jaspers and Steijlen, 2001). Similar to the MAC-like cases, irregular distribution of elastin fibres has been observed in individuals with Rombo syndrome, however these have been described as similar to 'swathes of steel wool' (Van Steensel, Jaspers and Steijlen, 2001). This is noticeably different from the 'ball-like' aggregations described in MAC-like cases (Schaller *et al.*, 2010), however similarities can be drawn in the inability for cells to regularly distribute elastin. More Rombo, MAC and MAC-like cases would need to be specifically observed for elastin distribution to identify whether these differences are distinct features that differentiate MAC from Rombo and MAC-like ductal proliferations.

4.3.4 Sweat duct morphology

Human sweat glands can be largely defined as apocrine or eccrine. Apocrine glands are involved in secretion into hair ducts; specialised apocrine glands include the ceruminous gland which produces cerumen (earwax) in the external auditory canal and the mammary gland (Cardoso and Calonje, 2015). Eccrine glands secrete directly to the skins surface and are expansive in their localisation in human skin in comparison to other mammals where they are mainly found on the palms and soles of the feet (Murota *et al.*, 2015). Both eccrine and apocrine sweat glands change in activity during puberty, particularly in their joint involvement in response to psychological stress (Harker, 2013). Eccrine glands show an increase in sweat secretion, induced by acetylcholine, at around the age of 12, particularly on the axilla, whereas apocrine glands remain inactive until puberty (Harker, 2013; Murota *et al.*, 2015).

Apocrine and eccrine sweat glands do share some morphological similarities. For instance, both glands are composed of a secretory unit and a ductal epithelium (Hibbs, 1958). However the ducts of eccrine

glands are long and thin, in comparison to the shorter and thicker ducts of the apocrine gland (Lu and Fuchs, 2014). In contrast, the larger gland region of the apocrine glands allows it to be positioned in the same region of the middle dermis to the smaller eccrine gland, in spite of their difference in duct length. The secretory region of eccrine glands constitutes an inner layer of secretory cells surrounded by myoepithelial cells, both of which are encircled by a basement membrane (Lu and Fuchs, 2014).

Traditionally, various malignant sweat gland tumours were classified by whether they originate from eccrine or apocrine glands (Cardoso and Calonje, 2015). MAC was originally reported as being of eccrine origin, due to the identification of intact eccrine glands and morphologically similar tumour cells (Goldstein, Barr and Santa Cruz, 1982). Immunohistochemical studies into MAC origin propose that the presence of S-100 and CF-1 positive cells is a suggestion of its eccrine origin (Ongenae *et al.*, 2001). However it has also been reported as having sebaceous differentiation and is now believed to have diverged from follicular apocrine glands (Cardoso and Calonje, 2015).

The contrasting evidence for the cellular origin of MAC may once again be explained by the range of clinical and histopathological features and diagnoses represented by MAC and MAC-like adnexal neoplasms. It could be the case that some clinical features grouped under the term MAC could be derived from multiple causes and cellular origins and should as such be treated as separate diseases. Some studies suggest that, as well as the more common differential diagnoses, syringoid eccrine carcinoma could be grouped with MAC (Cardoso and Calonje, 2015). As the name suggests, this eccrine tumour would stand apart from MAC in its histology; this may give weight to the argument that, instead of distinct clinicopathological entities, multiple malignant sweat gland tumours could be placed on a spectrum from the development of individual syringomas to a locally invasive neoplasm (Schaller *et al.*, 2010; Cardoso and Calonje, 2015).

The identification of one underlying genetic or environmental cause to these rare adnexal neoplasms could answer questions about how to categorise such diseases. Importantly, collating cases of the same genetic origin and describing the histopathological and clinical features will help to create a clear diagnostic and treatment path. It would also allow for larger combined datasets of individuals with a definitive diagnosis that is not subject to individual pathological review.

In this paper we describe five families with clinical and histopathological features that show similarities to MAC, Nicolau-Balus and Rombo syndromes (figure 4.1) and provide evidence that the described syndrome is a manifestation of germline variants in *MYH9.* Accordingly, we propose the term MALTA (MYH9 Associated eLasTin Aggregation) syndrome to reflect the underlying genetic basis of this disease.

4.3.5 Aims

This study gathered the sequencing data from five families with a rare but specific clinical and histopathological diagnosis. The aims were as follows:

1. To identify candidate variants linked to the development of irregular elastin aggregations and sweat duct proliferations.
2. To explore identified variants and identify any existing genotype-phenotype correlation with regards to this syndrome.
3. To describe a novel genetic syndrome associated with the development of sweat duct proliferations and irregular elastin aggregations.

4.4 Materials and methods

4.4.1 Study population

Families were identified by clinicians from collaborating centres following pathological review. Families 1 and 2 were recruited as part of the ethically approved Investigating Hereditary Cancer Predisposition (IHCAP) study (REC 12/EE/0478). Informed consent was obtained from all participants in accordance with institutional review board policies and local practices.

Three participants within two families (families 1 and 2) were selected for WES based on a diagnosis of sweat duct proliferations resembling MAC and irregular elastin aggregations as described by Schaller et al (Schaller *et al.*, 2010). Three additional cases from a third family were later recruited for WES analysis, with additional family members (including those unaffected with the disease) being used for further segregation analysis. Members of a further two Rombo families underwent Sanger sequencing for variants in the likely causal gene.

4.4.2 Clinical information

Family 1

Case 1 is a 44 year old male who first presented with swelling and hardening of the left temple region at the age of 19; an initial biopsy identified benign eccrine sweat duct proliferations which were classified as syringoadenoma (Schaller *et al.*, 2010). The individual has since presented swelling on both sides of the face with eczematous lesions on the neck and erythematous maculae on both elbows. Histopathological examination of biopsies taken from the facial lesions showed that ductal proliferations extended through the middle and deep dermis, infiltrating the connective tissue septa but not invading perineurally. Abnormal aggregations of elastic fibres were seen in additional biopsies. Several attempts were made to surgically remove the tumours, but these were unsuccessful and further surgical treatment was refused. This individual had no family history of skin tumours or cancer.

Family 2

Case 2-1 is a female who was diagnosed with Nicolau-Balus syndrome as a child after presenting with syringomata, milia and atrophodermia vermiculata (Schaller *et al.*, 2010). At age 26 she developed an indurated plaque on the right cheek, 6 cm in size, with an apparent punctum on its surface. Excision biopsy of the lesion showed it was deeply infiltrative and composed of sweat duct proliferations in a fibrous stroma. The proliferating ducts extended deep into the subcutis but, like case 1, did not display any perineural invasion. The mother of case 2-1 (case 2-2) similarly presented with atrophodermia vermiculata, syringomata on the neck and anterior chest with some infundibular facial cysts that showed evidence of ruptures and lesions on the buttocks. Biopsies from both cases showed the same abnormal, round aggregations of elastic fibres in the papillary dermis (figure 4.2A), however biopsies from case 2-2 did not show any ductal proliferation. There was no other family history of skin tumours or cancer.

Family 3

We identified three additional cases from one family (figure 4.3). Two are siblings (case 3-1 and case 3-2; male and female respectively) and the third is their male cousin once removed (case 3-3). Although these three individuals were diagnosed with Rombo syndrome and presented with atrophodermia vermiculata and further irregular elastic fibre distribution, the identification of sweat-duct proliferations during biopsies suggests that these are not classical Rombo cases. There are reported to be additional affected individuals across three generations, although the exact pathology of these individuals is unknown.

Family 4 and Family 5

A further six individuals from two additional families underwent targeted sequencing of the candidate gene. Family 4 were previously reported by Michaelsson et al in the first paper describing the clinical and histopathological features of Rombo syndrome (Michaelsson, Olsson and Westermark, 1981). The individual from family 5 had no family history of skin disorders but presented with atrophodermia vermiculata and multiple milia on the face as well as abnormal distribution of elastin. He was diagnosed with Rombo syndrome, with sweat duct proliferations identified during retrospective review of his histology (Van Steensel, Jaspers and Steijlen, 2001).

4.4.3 Germline whole exome sequencing and variant filtering

DNA from families 1 and 2 was extracted from blood and prepared for PE125 WES using the Nextera Rapid Capture Exome enrichment kit (Illumina). Sequencing was performed on Illumina HiSeq-4000 machines. Cases 3-1, 3-2 and 3-3 underwent PE100 WES on a HiSeq-2500 and were added to the dataset as FASTQ files. VCF files were generated using a standard pipeline following GATK best practice recommendations for whole exome data (see Chapter 2: Methods for further details).

Variants were filtered to select 732 rare (AF < 0.01 in 1000 genomes), protein-affecting variants (loss of function, predicted deleterious and damaging missense (by SIFT and PolyPhen respectively) and inframe indels). An in-house sequenced control set of 187 non-skin disorder related samples was used to select novel variants. Genes were selected that contain different variants within each family and that segregated with the phenotype where possible. None of the candidate variants summarised in this study have been identified in 1000 genomes, gnomAD or the in-house control set.

4.4.4 Validation and genotyping by Sanger sequencing

Candidate variants were validated by Sanger sequencing. Qubit dsDNA HS kit (Invitrogen) was used to quantify DNA. Custom flanking primers were designed for each variant and DNA fragments were amplified by PCR. The products were sequenced on an ABI Genetic Analyser (Applied Biosystems) using BDT V3.1 (Invitrogen) according to the manufacturer's instructions for Sanger sequencing.

The exons of candidate gene *MYH9* were analysed by Sanger sequencing in probands from families 4 and 5. Once a candidate variant was identified, genotyping was performed in other family members

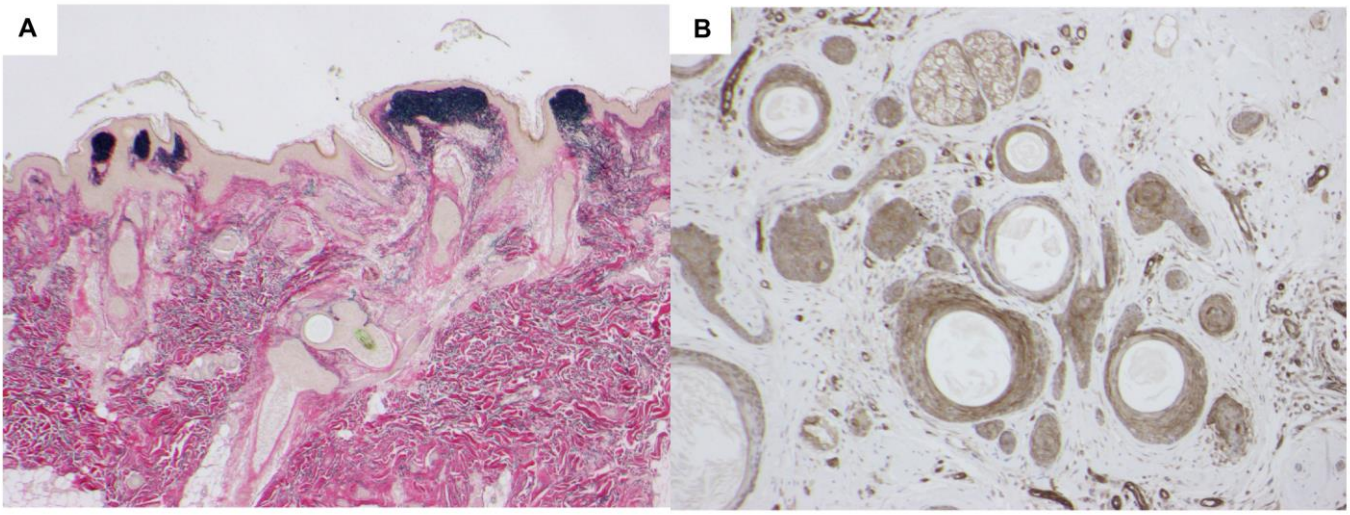


Figure 4.2: Elastin aggregation and MYH9 staining in MALTA cases. (A) irregular distribution of elastic fibres in case 2-1; (B) superficial keratocysts stained for MYH9 in case 1. The glandular component of this case's tumour showed a similar staining pattern

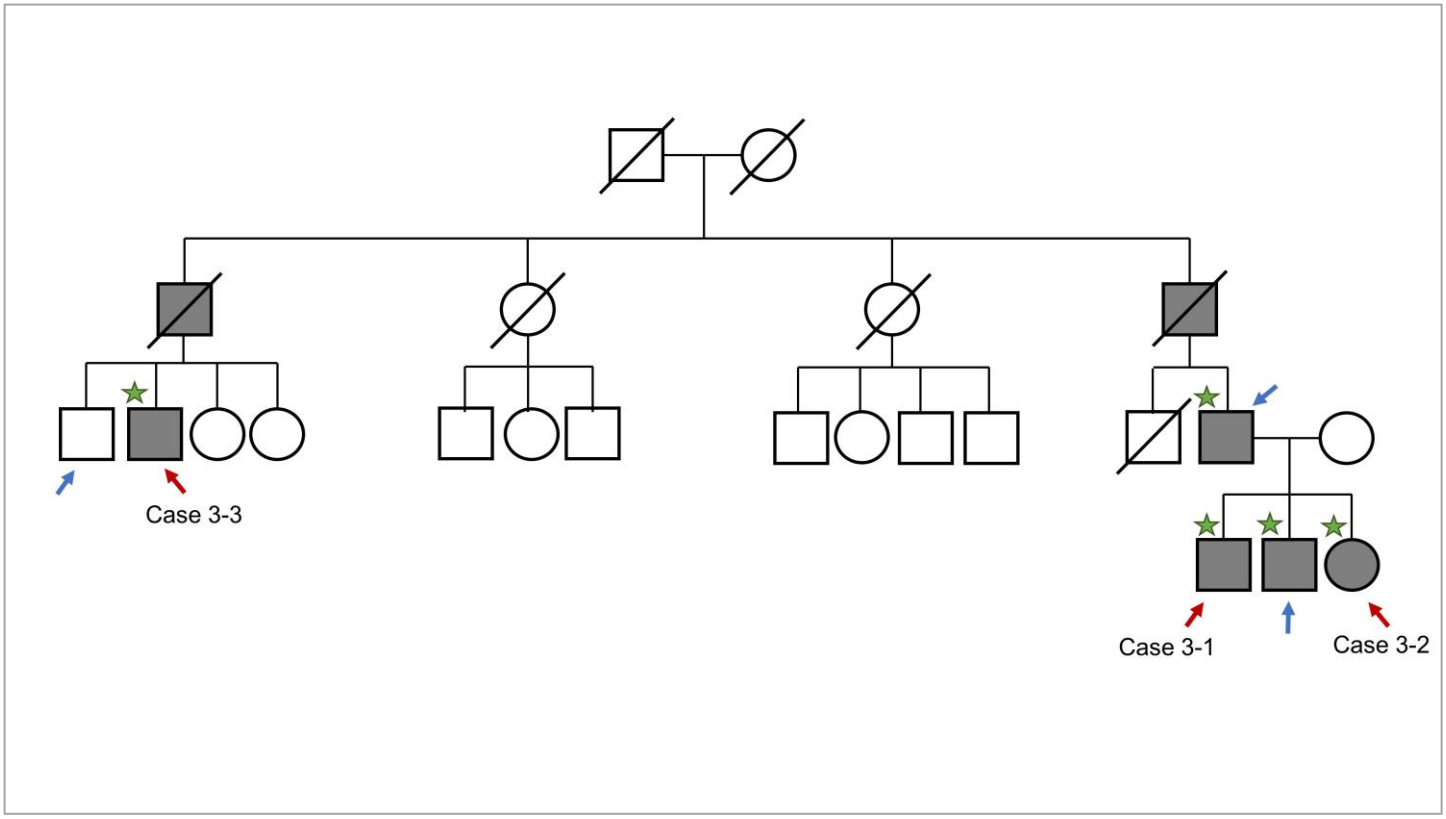


Figure 4.3: The pedigree for family 3. Red arrows indicate individuals that underwent WES. Blue arrows indicate individuals that underwent genotyping for the *MYH9* variant c.5492_5499delGGCTGCCT. Green stars indicate individuals who were heterozygous for this variant.

where available using Sanger sequencing to sequence the surrounding region. An additional three members of family 3 (one unaffected and two affected) and five members of family 4 (one unaffected and four affected) were genotyped using this method to allow for further segregation analysis.

4.4.5 Spatial analysis of *MYH9* variants

The protein structure of the *MYH9* product (non-muscle myosin IIA (NMMIIA)) was analysed using the mechanistic protein prediction tool Mechismo (Betts *et al.*, 2015), to assess variants in the context of key functional elements of the protein, including the ATP binding site. The NMMIIA amino acid substitutions identified in this study and their effect on protein interactions was studied. Each assessed variant generated a Mechismo score (sum of (1 + maximum absolute change in pair-potential)) to predict the effect it has protein-protein, protein-chemical and protein-DNA/RNA (Betts *et al.*, 2015).

With the aim of ascertaining how *MYH9* variants may produce multiple non-overlapping phenotypes, a comparable analysis was performed on the previously published pathogenic variants in *MYH9*. A set of 60 variants (Althaus and Greinacher, 2009; Saposnik *et al.*, 2014) described in the literature as causing MRPD, which may be coupled with renal impairment, hearing loss, and cataracts, were analysed for their effect on protein interactions.

4.4.6 Conservation of myosins at mutated regions

The amino acid sequences of myosin genes from the ten major classes were downloaded in the form of FASTA files and aligned using CLUSTALW to identify conserved regions. Myosin genes explored included class I *MYO1A* and *MYO1D*, class II *MYH1* and *MYH9*, class III *MYO3A*, class V *MYO5A*, class VI *MYO6*, class VII *MYO7A*, class IX *MYO9A*, and class X *MYO10*. Aligned files were viewed in BioEdit. Amino acid changes identified in MALTA samples and MRPD samples (Althaus and Greinacher, 2009) were analysed for conservation across different myosin classes. This was further expanded to look specifically at the class II conventional myosins encoded by the genes *MYH1, MYH3, MYH7, MYH9, MYH10, MYH11, MYH13, MYH14,* and *MYH15*.

4.4.7 Tumour immunohistochemistry

Tumour blocks were available for case 1, IHC was performed for NMMIIA using mouse monoclonal antibodies ab117572 and ab122978 (Abcam). Neither antibody targeted the mutated region identified within this individual.

4.4.8 Tumour whole exome sequencing and variant filtering

Tumour DNA for case 2-1 was extracted from cut scrolls of FFPE using an ultrasonicator (Covaris). Tumour and germline DNA (from blood) for this case were prepared for sequencing using Agilent SureSelect Human All Exon 50Mb kit following manufacturer's instructions and were sequenced on the Illumina HiSeq-4000 as previously described (Tarpey *et al.*, 2013). Raw sequencing data were analysed with a somatic variant calling pipeline using Mutect2 (GATK v3.7) to generate a VCF file of somatic variants (see Chapter 2: Methods for further details).

Variants were filtered to select those with a somatic variant allele frequency (VAF) of no less than 10%. Protein-affecting variants (loss of function, predicted deleterious and damaging missense (by SIFT and PolyPhen respectively), and inframe indels) were selected, creating a set of 245 somatic variants. These were manually searched for their occurrence in known oncogenes as well as in skin related cancers (which were histologically categorised into carcinoma, malignant melanoma, adnexal tumour or other) according to the COSMIC database (Forbes *et al.*, 2017).

All variants of interest were checked in IGV for coverage across the region and further validated using Sanger sequencing.

4.5 Results

4.5.1 Germline whole exome sequencing

The WES data from six individuals were analysed using an optimised in-house VCF generation pipeline to create a set of 732 different high confidence rare protein-affecting variants. Only one gene, *MYH9*, harboured protein-affecting variants in all families, with each variant appearing in all affected family members. The NMMIIA gene *MYH9* contained a missense variant (NM_002473:c.2012C>T, p.Cys671Tyr) in case 1 that was predicted to be deleterious and damaging to protein function. An inframe deletion (NM_002473: c.1767_1769delTCC, p.Met589_Asp590delinsIle) in cases 2-1 and 2-2 of family 2 and a frameshift deletion (NM_002473: c.5492_5499delGGCTGCCT, p.Gln1831Leufs*20) in cases 3-1, 3-2 and 3-3 (figure 4.3) were also identified in this gene. None of the three variants had previously been observed in phase 3 1000 genomes (Auton *et al.*, 2015) or gnomAD (Lek *et al.*, 2016). Three additional members from family 3 were recruited for genotyping, including two individuals that were affected with the described phenotype and a third unaffected individual. The identified frameshift deletion segregated completely with the MALTA phenotype within this family (figure 4.3). No unaffected members of families 1 or 2 were available for segregation analysis.

Upon the identification of a candidate gene for MALTA predisposition, two additional families underwent targeted sequencing of the *MYH9* gene. Both families were diagnosed with Rombo syndrome and presented with irregular distribution of elastin fibres. A missense variant (NM_002473:c.706G>T, p.Gly236Cys) was identified in family 4 and an inframe indel (NM_002473:c.694_695delinsAA, p.Ser232Asn) was observed in family 5. Both variants result in a one amino acid change which is predicted to be deleterious to protein function and neither have been previously observed in phase 3 1000 genomes (Auton *et al.*, 2015) or gnomAD (Lek *et al.*, 2016). Additional members of family 4 were recruited for segregation analysis, with the missense variant fully segregating with phenotype. All identified *MYH9* variants are described in table 4.1.

4.5.2 Spatial analysis of *MYH9* variants

Variants were mapped to the predicted NMMIIA protein structure generated by Mechismo. Four of the five identified variants clustered in the myosin head (figure 4.4). It was predicted that the variant c.694_695delinsAA identified in family 5 affects the binding of ADP to myosin and generated a Mechismo score of 1, where scores of 1 or greater indicate a disruption to protein interactions. The generated model also indicates that the missense variants c.2012C>T and c.706G>T also lie within the ATP binding domain (figure 4.4), although neither are predicted to effect protein interactions by Mechismo.

Family	Proband Gender (age at diagnosis)	Diagnosis of Proband	Dermatological features in proband				Proband platelet count and volume	Family members sequenced	*MYH9* Variant	Amino Acid change	VEP Consequence	Segregation	Variant identification
			MAC-like ductal proliferations	Irregular elastic fibre distribution	Atrophodermia vermiculata	Syringomas							
1	Male (9)	MAC-like ductal proliferations	Yes	Yes	Yes	Yes	Normal	1	c.2012C>T	p.Cys671Tyr	Missense variant	Sporadic	WES
2	Female (9)	Nicolau-Balus syndrome/MAC-like ductal proliferations	Yes	Yes	Yes	Yes	Normal	2	c.1767_1769delTCC	p.Met589_Asp590delinsIle	Inframe deletion	Not available	WES
3	Male (17)	Rombo syndrome	Yes	Yes	Yes	Yes	Normal	6	c.5492_5499delGGCTGCCT	p.Gln1831Leufs*20	Frameshift deletion	Yes	WES
4	Male (7)	Rombo syndrome	No	Yes	Yes	Not described	Not described	5	c.706G>T	p.Gly236Cys	Missense variant	Yes	Targeted sequencing
5	Male (24)	Rombo syndrome	Yes	Yes	Yes	No	Unknown	1	c.694_695delinsAA	p.Ser232Asn	Inframe indel	Sporadic	Targeted sequencing

Table 4.1: MALTA Families analysed and *MYH9* variants identified in each family.

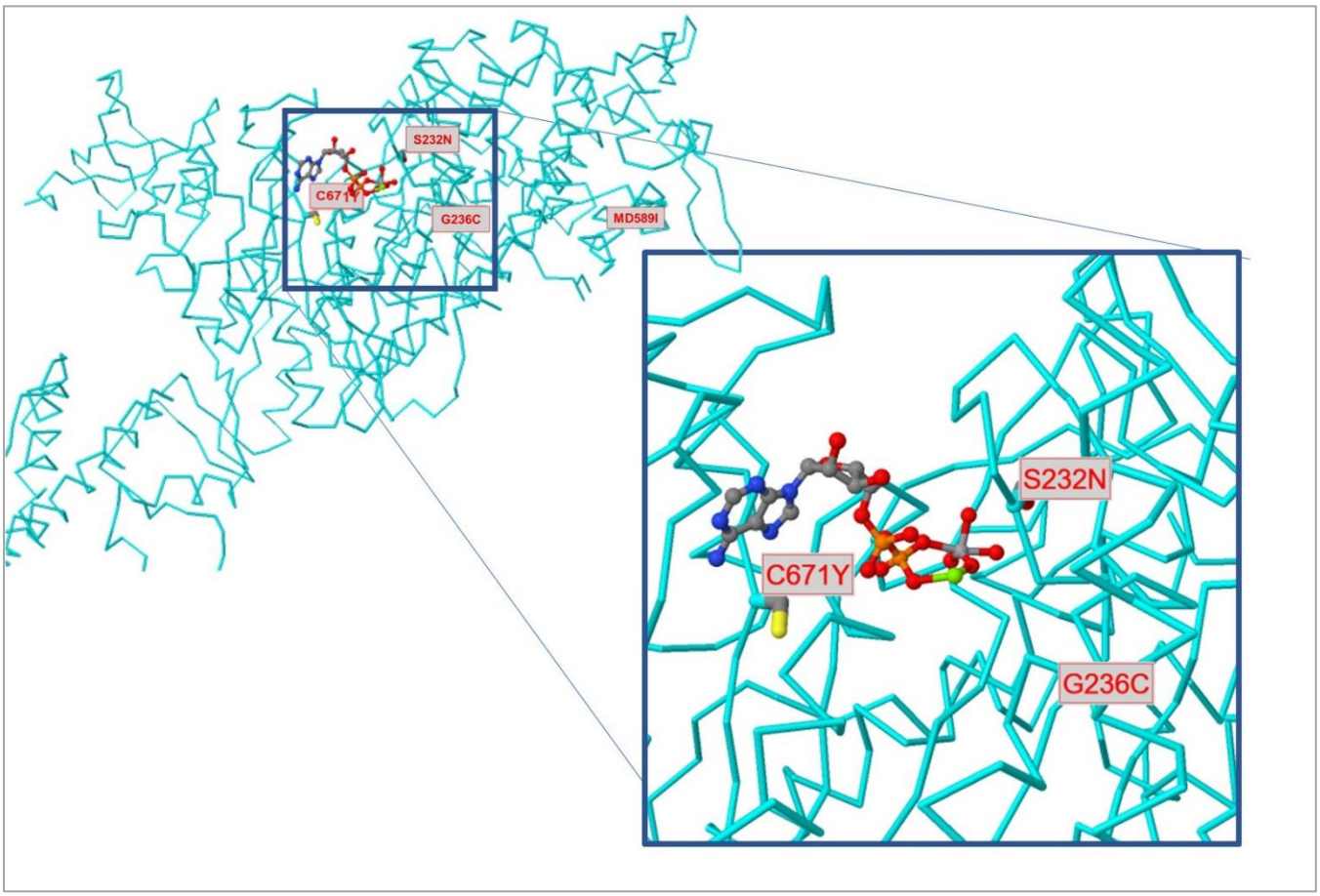


Figure 4.4: Identified *MYH9* variants mapped to the head domain of non-muscle myosin IIa. Three of the identified variants fall within the ATP binding pocket (inset).

In contrast to other motor domain variants identified, the inframe deletion c.1767_1769delTCC appears, according to the 3D protein structure, to affect the actin binding region. The frameshift variant identified is positioned at the end of the tail domain, resulting in the loss of key acetylation and phosphorylation sites.

To further understand how variants in *MYH9* could be creating two distinct phenotypes, variants that have been known to be linked to MRPD were also analysed using Mechismo to predict their effect on protein interactions. In contrast to variants detected in our study, most of which were located in the ATP-binding region of the myosin head, most of the MRPD-associated variants were located in myosin tail (figure 4.5). Importantly, unlike the variants identified in the MALTA syndrome families, the MRPD variants identified in the myosin head did not lie in either the ATP or actin binding sites which are key to myosin function. Six of the 60 MRPD variants that were analysed were predicted with high confidence to be within the binding regions for S100 calcium binding proteins A4, A5 or P. Only one of these variants, p.Arg1933* was predicted to interrupt the interaction. Another variant, p.Lys850Glu was predicted to affect the formation of both heterodimers with NMMIIB (*MYH10* gene product) and homodimers.

4.5.3 Conservation of myosin classes

The conservation of NMMIIA was observed across eight other myosin classes and specifically within class II (containing NMMIIA, B, and C). Amino acid regions that were substituted in the MALTA phenotype were shown to be highly conserved across all classes of myosin (table 4.2), with ten different myosins showing the same amino acid at all affected sites. The frameshift variant identified in the myosin tail is the exception to this, with three other myosins having shorter tail regions that would not encompass this variant and others not being conserved.

In comparison, seven amino acids that are substituted in MRPD did not show conservation across the ten tested myosins. No tested amino acid showed complete conservation across tested myosins from all classes. These regions were conserved within the nine tested class II myosins, with four sites showing complete conservation (table 4.2).

4.5.4 Tumour immunohistochemistry

Two tumours from case 1 were stained for NMMIIA. The variant identified in case 1 was a heterozygous missense and *MYH9* was expressed in the tumours (figure 4.2B). Staining in the tumour, in healthy skin and in other benign adnexal lesions was ubiquitous.

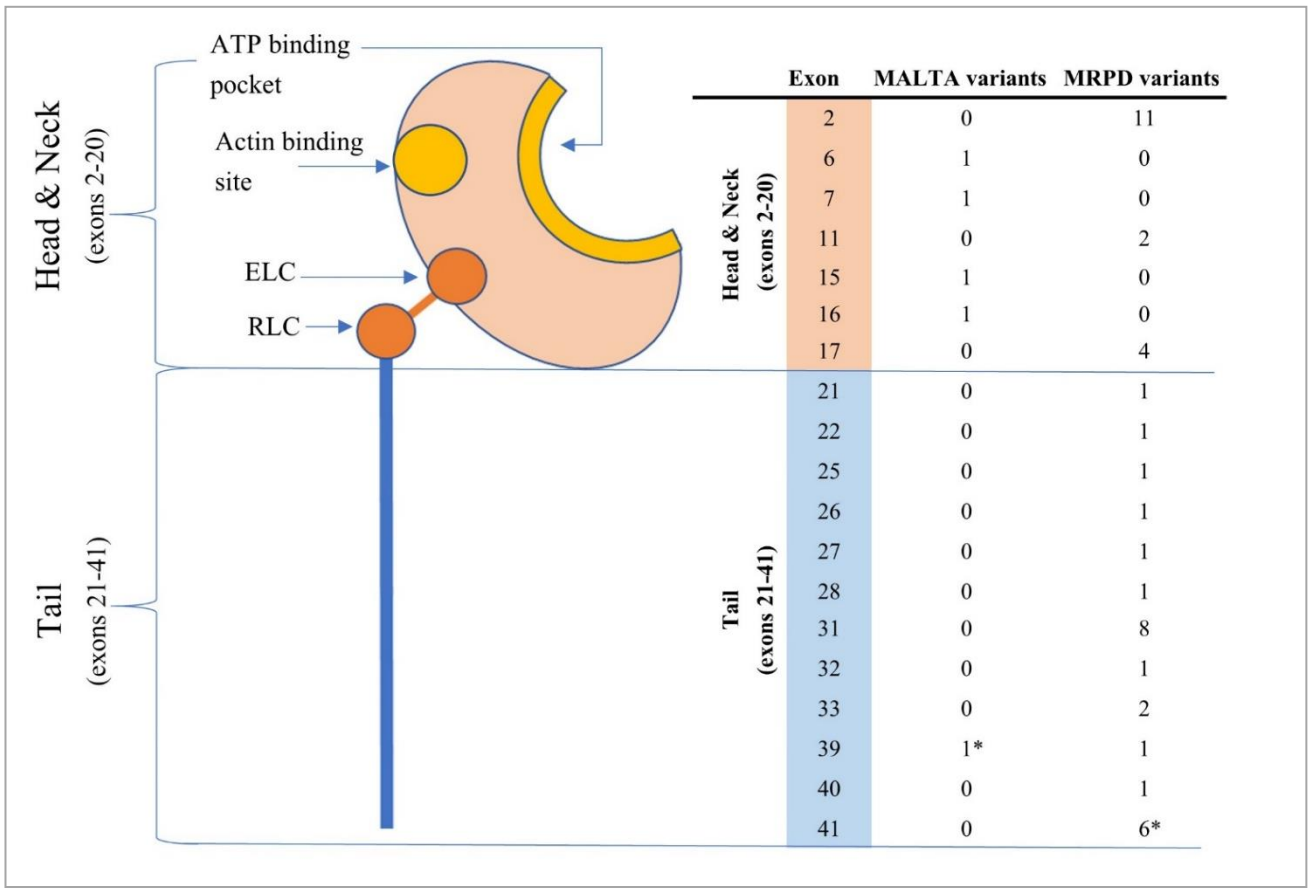
4.5.5 Tumour whole exome sequencing

Somatic variants identified in case 2-1 were analysed for known oncogenic genes and variants. One identified candidate variant was a heterozygous missense variant in the receptor tyrosine kinase proto-oncogene RET (c.2269G>A, p.Val757Met). This variant has not been previously described in the COSMIC database, however variants described in COSMIC within this exon and the surrounding exons

have been identified in three facial adnexal sebaceous adenoma tumours (including one sample diagnosed with Muir-Torre syndrome).

BAM files were manually searched for loss of heterozygosity of the identified *MYH9* variant. The variant, c.1767_1769delTCC, appeared to remain heterozygous within the tumour. There were no other protein-affecting variants seen in *MYH9* within the tumour sample.

The low DNA quality of the sequenced tumour sample (as confirmed by checking BAM files manually) meant any calculated mutation rate is likely to be highly inaccurate. However the general low rate of variants seen in this tumour sample (8.2 per Mb) is not representative of highly UV damaged DNA which can have a mutation rate upwards of 20 variants per Mb, suggesting that these lesions are not linked to UV damage (Mar *et al.*, 2013).



	Exon	MALTA variants	MRPD variants
Head & Neck (exons 2-20)	2	0	11
	6	1	0
	7	1	0
	11	0	2
	15	1	0
	16	1	0
	17	0	4
Tail (exons 21-41)	21	0	1
	22	0	1
	25	0	1
	26	0	1
	27	0	1
	28	0	1
	31	0	8
	32	0	1
	33	0	2
	39	1*	1
	40	0	1
	41	0	6*

Figure 4.5: MALTA variants identified within this study and *MYH9*-related platelet disorder variants were mapped to the exons of *MYH9*. All variants are missense or inframe aside from those indicated by * which are protein truncating. When placed into the context of the non-muscle myosin structure, the differential distribution of variants according to disease can be seen across the head and tail regions. ELC: essential light chain, RLC: regulatory light chain

Myosin	*Amino acid position*	*96*	*232*	*236*	*371*	*589*	*590*	*671*	*702*	*1165*	*1424*	*1831*	*1841*	*1933*
Class I	MYO1A	S	S	G	V	N	D	C	R	~	~	~	~	~
	MYO1D	E	S	G	V	K	D	C	R	~	~	~	~	~
Class II	MYH1	A	S	G	A	K	D	C	R	R	D	V	E	~
	MYH3	A	S	G	A	K	D	C	R	R	D	T	E	E
	MYH7	A	S	G	A	K	D	C	R	R	D	A	E	~
	MYH9	S	S	G	A	M	D	C	R	R	D	Q	E	R
	MYH10	S	S	G	A	M	D	C	R	R	D	A	E	R
	MYH11	S	S	G	A	M	D	C	R	R	D	Q	D	R
	MYH13	A	S	G	A	K	D	C	R	R	D	A	E	~
	MYH14	S	S	G	A	M	D	C	R	R	D	I	E	V
	MYH15	S	S	G	A	K	D	C	R	Q	D	A	E	~
Class III	MYO3A	T	S	G	L	R	D	C	R	~	~	Q	~	~
Class V	MYO5A	A	S	G	L	K	D	C	R	L	G	G	~	~
Class VI	MYO6	T	S	G	L	N	D	C	D	~	~	~	~	~
Class VII	MYO7A	G	S	G	L	R	D	C	R	I	P	I	L	T
Class IX	MYO9A	T	S	G	L	T	D	C	R	Q	Q	I	E	G
Class X	MYO10	S	S	G	L	R	D	C	R	Q	I	K	E	Q
Variants	MYH9	L	N	C	N		I		Y	C/H	C	N/H	fs	K	X
			**	**	*			**	**	*		*			

Table 4.2: Conservation of amino acid sites shown to be substituted or truncated in MALTA (grey) and MRPD (white) across eight classes of myosins, focusing on class II which contains MYH9 (red) and the other non-muscle myosins (blue). ** site conserved across all tested classes, * site conserved across class II myosins.

4.6 Discussion

All five MALTA families analysed within this study carried variants in the NMMIIA gene *MYH9*. There are 15 different class II myosins, three of which do not appear in muscle cells and are so called non-muscle myosins (NMMs). These are encoded by the genes *MYH9* (NMMIIA), *MYH10* (NMMIIB), and *MYH14* (NMMIIC). The three NMMs are expressed in different cell types, with NMMIIA and NMMIIB being expressed in epithelial and endothelial cells and NMMIIA being expressed in circulating platelets (Betapudi, 2014).

NMMs have previously been linked to a diverse range of cellular functions including cell migration and invasion, protein and organelle localization, and cell signalling (Ouderkirk and Krendel, 2014), making them key interactors in many of the hallmarks of cancer. Pathogenic variants in the unconventional myosin *MYO1A* have been previously seen in a high proportion of colorectal and gastric tumours (Mazzolini *et al.*, 2012, 2013) and the gene has since been labelled a tumour suppressor gene (Ouderkirk and Krendel, 2014). Similarly the myosin XVIIIB gene *MYO18B* has been classed as a candidate tumour suppressor gene and has been shown to be deleted or carrying somatic pathogenic mutations in lung, ovarian and colorectal cancers (Nishioka *et al.*, 2002; Yanaihara *et al.*, 2004; Nakano *et al.*, 2005). The NMM gene *MYH9* has been described as a tumour suppressor gene in SCCs and functional work has identified a specific role for NMMIIA in p53 localisation and stabilisation (Schramek *et al.*, 2014).

Within this study, all tested individuals with the MALTA phenotype, including a combination of MAC-like sweat duct proliferations, irregular distribution of elastin, atrophodermia vermiculata, and syringomas, carried a novel protein-affecting *MYH9* variant. Families 1 and 5 contained just one sporadic individual, with no available family members for further testing. Families 2, 3 and 4 had strong histories of MALTA histopathologic features, allowing multiple affected, and in some cases unaffected, individuals to be genotyped, with variants segregating fully with phenotype. Based on the current data, the identified *MYH9* variants appear to be fully penetrant.

Variants in *MYH9* have been previously associated with MRPD. No participants identified in this study exhibited any of the features of MRPD, with all individuals displaying normal platelet size and count and no hearing impairment. Similarly, publications on the characteristics of MRPD suggest no skin abnormalities (Althaus and Greinacher, 2009; Saposnik *et al.*, 2014). This suggests that *MYH9* is producing two distinct phenotypes. The majority of MRPD variants fall in the tail region of NMMIIA (25 variants out of 42) (figure 4.5). In comparison, 4 of the 5 MALTA variants identified in this study lie within the head region, which contains both the ATP and actin binding regions. Only one exon, exon 39 in the tail region, contains a variant in both MALTA and MRPD. The exon 39 MALTA variant is a frameshift variant which is predicted to cause an early stop codon. In contrast, loss of function variants in MRPD are only seen in the final exon of the gene.

By mapping the identified variants onto the predicted 3D protein structure of NMMIIA, four of the five *MYH9* variants can be seen in the head region, with three clustering in the ATP binding pocket. Only one of the three variants, p.Ser232Asn (c.694_695delinsAA), was predicted to interrupt the ATP/ADP interaction, however the close proximity of the other two missense variants to the binding pocket may also suggest they are causing a deficiency in ATP binding or in the resulting conformational change that is key to the motor function (Betapudi, 2014).

To further study the two distinct phenotypes produced by *MYH9* variants, MRPD variants were also plotted into Mechismo to identify any disruption to protein interactions. A number of the variants input were predicted to interrupt the interaction with S100 calcium binding proteins A4, A5 and S100P. These proteins are required for controlling myosin II phosphorylation and filament assembly (Clark *et al.*, 2008). This corroborates previous findings that the *MYH9* variants described in MRPD disrupt filament assembly by affecting the α-helical coiled-coil tail structure (Kriajevska *et al.*, 2000). Additionally, variants were predicted to interrupt binding of NMMIIA to NMMIIB and to itself thereby preventing the formation of hetero and homodimers.

By looking at the conservation of myosins, it was observed that sites with variants in the MALTA phenotype were conserved across all tested eight myosin classes. This suggests that the *MYH9* variants identified within this study interrupt a function that is integral to all myosins such as ATP binding. Whereas MRPD variants were in regions that are only conserved in the class II myosins, also known as conventional myosins, which encompasses the three NMMs NMMIIA, NMMIIB, and NMMIIC. We can hypothesise that, in contrast to the MALTA *MYH9* variants, the MRPD variants interrupt a function that is not integral to all myosins, but key to class II myosins specifically. Such class II specific functions include the formation of intracellular hexameric monomers, which was suggested to be interrupted by MRPD variants in Mechismo.

The MRPD phenotype extends beyond platelet abnormalities to affect a wide range of tissues and cell types including leukocytes, the kidney and the cochlea. It remains unknown how the disruption of *MYH9*, a gene that is so essential to cell function in many human tissues, leads to very specific phenotypes such as deafness and cataracts. One study found that expression of combinations of the three NMM isoforms in a tissue specific manner drives differential myosin functions, showing evidence of partial co-localization of the two proteins suggesting some overlap in function in addition to some unique cellular roles (Marini *et al.*, 2006).

The expression of more than one NMM in most cell types suggests that these proteins play similar if not identical roles in these cells. In platelets, only NMMIIA is expressed which likely explains the phenotype in MRPD. It is possible that within the areas of the dermis affected by MALTA syndrome, NMMIIA plays a unique role that cannot be recovered by the presence of another NMM. Within skin fibroblasts, NMMIIA expression was shown to increase during dermal wound healing and scar tissue

remodelling, while expression of NMMIIB remains stable, suggesting differential roles for the two proteins (Bond *et al.*, 2011). During this healing process, the extra-cellular matrix (ECM) components including elastin are secreted by fibroblasts and the elasticity of the ECM plays a role in the regulation of NMMIIA and NMMIIB during scar formation and fibroblast migration (Bond *et al.*, 2011; Almine, Wise and Weiss, 2012). It could be through this damage driven signalling cascade that *MYH9* variants in MALTA syndrome are affecting elastin secretion and distribution.

Loss of *MYH9* has been previously studied by knockouts and RNAi in the context of tumour development in head and neck SCC (Schramek *et al.*, 2014). Human and mice keratinocytes were used to further understand *MYH9* function and identified that as well as its well characterised role as a motor protein with actin, it played a key role in stabilizing p53 (Schramek *et al.*, 2014). Blebbistatin was used as a NMMIIA inhibitor to block the ATPase activity and show that p53 did not accumulate in the nucleus after DNA damage as it does without blebbistatin treatment (Schramek *et al.*, 2014). This suggests that *MYH9* could play an important role in the DNA damage response, and the link between MALTA proliferations and UV damage could be the result of a lack of this damage response.

A missense variant was identified in the tumour of case 2-1 in the proto-oncogene *RET*. Single nucleotide variants in *RET* are typically associated with multiple endocrine neoplasia and medullary thyroid cancers (Plaza-Menacho, Mologni and McDonald, 2014). However a number of other cancers including papillary thyroid cancer, non-small cell lung cancer, breast cancer, pancreatic cancer, and prostate cancer have all been associated with somatic *RET* fusions or overexpression of the gene (Plaza-Menacho, Mologni and McDonald, 2014). Within COSMIC, skin tumours were analysed for *RET* variants, with particular emphasis on facial adnexal tumours of which 15% of those tested carried somatic single nucleotide variants in *RET*. Other skin cancers with *RET* single nucleotide variants included malignant desmoplastic melanomas of the head, neck and arm. The identification of *RET* variants in a MALTA tumour may indicate that tyrosine kinase inhibitors could be explored as a new treatment option for affected individuals.

4.7 Summary

The in-depth analysis of MALTA individuals has provided a new insight into the genetic landscape of sweat duct neoplasms. Identifying and understanding these genetic factors has a strong impact on diagnosis and disease management. The rarity of such diseases has previously lead to a lack of research into predisposition and oncogenesis. Grouping together case studies that were previously diagnosed with a range of diseases under one genetic syndrome will help to provide studies into the wider group of *MYH9*-related sweat duct tumours. This study also shines a light on NMMs which may play a closer role in DNA damage response and cancer predisposition than previously thought.

5 Exploring the genetic landscape of early onset Adrenocortical carcinoma.

5.1 Introductory statement

The adrenocortical carcinoma patients described in this work were identified and recruited by Dr Ruth Casey who also provided extensive clinical information. Tumour immunohistochemistry analysis for available tumours was performed by Dr Alison Marker. Library preparation and sequencing of the germline DNA samples described in this chapter was performed by James Redman and the Department of Medical Genetics Stratified Medicine Core Laboratory. Sequencing data were downloaded and processed by me using an in-house WES pipeline generated by Alexey Larionov. The results published here are in part based on data generated by The Cancer Genome Atlas (TCGA), managed by the National Cancer Institute and the National Human Genome Research Institute. Controlled access data were requested and downloaded for the TCGA-ACC dataset. The RNA-sequencing data from TCGA were analysed using an RNA-sequencing analysis pipeline generated by myself. All downstream data analysis was designed and performed by me.

In recent years, there has been an increased focus on the histological classification of ACC tumours. The oncocytic variant of ACC is now recognised as a distinct subtype and shall herein be referred to as 'oncocytic ACC'. Another type of ACC, often called normal ACC or just ACC itself, shall be referred to as 'usual type ACC' in accordance with current terminology used in the literature and to distinguish it from all ACCs. Where neither an oncocytic nor usual type is specified, it should be assumed that text is referring to ACC as a whole.

5.2 Abstract

Adrenocortical carcinoma (ACC) is an endocrine cancer with a poor prognosis as it often presents at an advanced stage. Li Fraumeni syndrome, caused by germline pathogenic variants in the tumour suppressor gene *TP53,* is associated with a predisposition to adrenal tumours, particularly childhood cases. However, the genetic factors associated with ACC development outside the realms of *TP53* are largely unknown. The oncocytic subtype of ACC is comprised of oncocyte cells which can be defined by an increased number of mitochondria causing a granular cytoplasm. These tumours have been previously described as having a low malignant potential, however prospective analysis has shown that this distinct histological subtype can be malignant and aggressive. This study describes the germline whole exome sequencing of eight oncocytic and usual type ACCs. Gene interaction network analysis was used to identify clusters of related candidate genes. Additional analysis was completed to identify candidate variants within genes that interact with *TP53.* TCGA ACC germline data were examined for variants in any of the identified candidate genes. Additionally, RNA-sequencing data from TCGA were used to perform differential expression analysis to identify genes that may influence the development of an oncocytic phenotype in comparison to usual type ACC. MAPK pathway related G-protein coupled receptor genes whose products interact physically with *TP53* were identified with protein-affecting variants within the described set and

within the TCGA-ACC set and could influence ACC risk. RNA sequencing analysis revealed that cAMP-PKA pathway genes had a change of gene expression in the oncocytic subtype of the TCGA-ACC data. This work has identified potential new pathways to be explored in both ACC predisposition and oncocytic ACC development.

5.3 Introduction

5.3.1 Adrenocortical carcinoma

Adrenocortical carcinoma (ACC) is an extremely rare endocrine tumour with variable malignant potential. The incidence of ACC is predicted to be around 2 per million people per year globally and the disease accounts for around 0.2% of cancer deaths globally (Schteingart *et al.*, 2005).

The prognosis of ACC is generally poor as patients typically present with advanced disease (Bilimoria *et al.*, 2008). For patients with localised disease, surgical resection is the only potentially curative treatment, however recurrence after resection is common and management strategies are not well defined (Gonzalez *et al.*, 2007). Within one study, distant metastases were identified in 66% of patients during follow-up after surgery (Ayala-Ramirez *et al.*, 2013). Early diagnosis in ACC facilitates complete surgical resection and offers the best chance for prolonged disease free survival (Xiao *et al.*, 2015). This emphasises the need to understand the genetic landscape of ACC, allowing individuals at greater risk to undergo personalised surveillance strategies and enable earlier diagnoses.

5.3.2 Epidemiology of adrenocortical carcinoma

The occurrences of ACC are well documented in the literature. One American study notes not just sex and race as patient characteristics but also median income and level of education received (Bilimoria *et al.*, 2008). Several studies have noted a preponderance for the disease amongst women, who account for between 55-67% of cases (Gonzalez *et al.*, 2007; Bilimoria *et al.*, 2008; Fassnacht, Kroiss and Allolio, 2013). The average age of diagnosis varies, and studies indicate that peak the incidence is between 40-55 years of age (Bilimoria *et al.*, 2008; Fassnacht, Kroiss and Allolio, 2013). Paediatric cases of ACCs have been described although these are extremely rare and are diagnosed with an incidence of 0.3 to 0.4 per million children (Michalkiewicz *et al.*, 2004). Cases of ACC in children are much higher in a south Brazilian population due to a high prevalence of the p.Arg337His allele of the *TP53* tumour suppressor gene (Ribeiro *et al.*, 2001; Wasserman *et al.*, 2015). It has since been suggested that between 50 and 80% of childhood occurrences of ACC are associated with pathogenic variants in *TP53* (Gonzalez *et al.*, 2009; Wasserman *et al.*, 2015). A younger age of cancer diagnosis is often associated with genetic predisposition factors. This has also been seen in ACC, with ~51% of cases diagnosed below the age of 20 being *TP53* pathogenic variant carriers (35 carriers out of 69 cases) in comparison to ~4% of cases diagnosed above the age of 20 (7 carriers out of 157 cases) (Herrmann *et al.*, 2012; Raymond, Else, *et al.*, 2013; Wasserman *et al.*, 2015).

5.3.3 Clinical Features of adrenocortical carcinoma

The large size, higher density, and heterogeneous enhancement of ACC tumours on computed tomography (CT) imaging often allows them to be distinguished from smaller adenomas. Tumours present bilaterally in 1-10% of cases (Latronico and Chrousos, 1997; Bilimoria *et al.*, 2008; Else *et al.*, 2014), with one large American study being on the lower end of this range with 1.1% of cases having bilateral tumours upon diagnosis (Bilimoria *et al.*, 2008). The description of ACC samples within TCGA

identified 56% (51 out of 91) of tumours involving the left gland (Zheng *et al.*, 2016). ACCs have been shown to commonly metastasise to the liver, lung, and bone, with the liver and lung each corresponding to around 10% of distant metastasis identified (Bilimoria *et al.*, 2008).

Approximately 50% of ACCs can be described as clinically functioning or hormone secreting (Schteingart *et al.*, 2005); meaning the tumours are autonomously secreting adrenal hormones including cortisol and androgens. The biochemical secretory pattern of a tumour is important for diagnosis and 40-60% of patients first present with hormonal excess (Luton *et al.*, 1990; Allolio and Fassnacht, 2006; Fassnacht, Kroiss and Allolio, 2013), of which 50-80% present with hypercortisolism (Else *et al.*, 2014). Presenting symptoms can include diabetes mellitus, osteoporosis, and hypokalemia or hypertension as a result of glucocorticoid-mediated mineralocorticoid receptor activation (Else *et al.*, 2014). Other commonly produced hormones reported by Gonzalez et al include androgens in 17% (12 out of 72 cases), aldosterone in 7% (5 out of 72 cases), estrogen in 1% (1 out of 72 cases), and mixed hormone production in 21% of tumours (15 out of 72 cases) (Gonzalez *et al.*, 2007). Adrenal androgen secretion has been identified in larger proportions of hormone secreting ACCs in other studies, with the percentage of secreting tumours reaching up to 60% (Else *et al.*, 2014). Symptoms of excess androgen secretion include male pattern baldness, hirsutism, virilization, and menstrual irregularities (Else *et al.*, 2014).

For tumours that are not hormone secreting, disease presentation is often due to 'mass effect' symptoms, for example abdominal or flank pain due to large tumour size (Bharwani *et al.*, 2011). For this reason, patients with non-functioning tumours are more likely to present at a late stage which could be a causal factor in the ~20-30% of cases who present with distant metastases (Gonzalez *et al.*, 2007; Bilimoria *et al.*, 2008; Else *et al.*, 2014). However, cortisol producing ACCs have been shown to have a reduced overall survival in comparison to non-functioning and androgen-producing tumours (Gonzalez *et al.*, 2007).

5.3.4 Histological subtypes

Of the non-usual type ACCs, the most common histological subtype is the oncocytic variant, which is predominantly composed of oncocytes (Else *et al.*, 2014). These oncocytes can be defined an increased number of mitochondria causing the epithelial cells to have a granular cytoplasm. As of 2010 only 17 cases of oncocytic ACC had been described in the literature (el-Naggar, Evans and Mackay, 1991; Alexander and Paulose, 1998; Kurek *et al.*, 2001; Hoang, Ayala and Albores-Saavedra, 2002; Seo, Henley and Min, 2002; Bisceglia *et al.*, 2004; Sang *et al.*, 2004; Tanaka *et al.*, 2004). Although oncocytic tumours have been suggested to act less aggressively than usual type ACCs, distant metastases have been identified in a number of cases (Tanaka *et al.*, 2004). This disparity may be a reflection of the small number of described cases leading to an inaccurate picture of the aggressiveness of the disease; for example, the oncocytic ACCs identified within this study were malignant and more aggressive than expected from previous clinical case studies. Additionally, diagnosing the malignant potential of these tumours has been noted as difficult in a number of studies due to the solid growth pattern of oncocytic tumours (Kurek *et*

al., 2001; Sang *et al.*, 2004; Else *et al.*, 2014), creating challenges for disease management (De Krijger and Papathomas, 2012).

The myxoid tumour variant of ACC is named due to their abundant production of extracellular myxoid material (Else *et al.*, 2014). This rare subtype has been described a number of times in the literature, both in a benign setting and with distant metastasis (Brown *et al.*, 2000). Interestingly one study described a myxoid case which had a distinct gene expression pattern, separating it from other high-grade ACCs (Giordano *et al.*, 2003). However, the rarity of this subtype means that this finding has not been replicated.

The sarcomatoid subtype is used to describe tumours with sarcomatous and carcinomatous features (Coli *et al.*, 2010). This tumour type has been diagnosed between the ages of 29 and 79, and diagnosis is often made postoperatively (Coli *et al.*, 2010). Behaviour of these tumours is often extremely aggressive and usually portends to a poor prognosis (Coli *et al.*, 2010; Else *et al.*, 2014).

5.3.5 Treatment

For non-metastatic ACC patients, surgery is widely accepted as the most curative therapeutic approach (Icard *et al.*, 1992; Allolio and Fassnacht, 2006; Fassnacht, Kroiss and Allolio, 2013), and is the most commonly used strategy, either alone or in combination with chemotherapy or radiotherapy (Gonzalez *et al.*, 2007; Bilimoria *et al.*, 2008). For individuals with hormone excess producing tumours, the levels of cortisol need to be closely monitored and controlled prior to surgery as elevated cortisol levels have been associated with poor wound healing and infection (Else *et al.*, 2014).

As well as being the primary therapeutic agent for ACCs, mitotane can be used to control cortisol levels in hormone-excess tumours by increasing mitochondrial function and therefore affecting the extra-adrenal metabolism of cortisol and androgens (Luton *et al.*, 1990; Latronico and Chrousos, 1997). Mitotane was developed from the insecticide dichlorodiphenyltrichloroethane which was described with adrenolytic activity in dogs in 1948 (Nelson and Woodard, 1948). An isomer of this compound was isolated and used by Bergenstal et al in 1960, showing patient response to this type of therapy (Bergenstal *et al.*, 1960). Mitotane is still considered the most effective chemotherapeutic agent for ACCs despite its low response rate (Berruti *et al.*, 2005; Gonzalez *et al.*, 2007; Bilimoria *et al.*, 2008). It is more commonly used as an adjuvant therapy alongside radical resection and in these contexts has been shown to have a significant effect on recurrence-free survival (Terzolo *et al.*, 2007). Mitotane has also been used in combination with chemotherapy regimens including cisplatin and etoposide, and cisplatin, etoposide and doxorubicin. Within one study, ~19% of patients (13 out of 67) treated with mitotane had a complete or partial response to treatment. Although, those who responded to mitotane had been treated with this alone and none of the patients treated with additional chemotherapy had shown a response (Gonzalez *et al.*, 2007).

5.3.6 Genetics of adrenocortical carcinoma

Li Fraumeni syndrome is a well-studied, highly penetrant, cancer predisposition syndrome that confers an increased risk to a number of early onset cancers. It was first described in 1961 by Li and Fraumeni,

whose analysis of families with childhood rhabdomyosarcoma identified cases of sarcoma, leukaemia, brain tumours, breast cancer, and ACC (Li and Fraumeni, 1969). Amongst the 44 early onset cancer cases (diagnosis at less than 15 years of age) identified in this study of four families, 9% had adrenal tumours (Li and Fraumeni, 1969). Li Fraumeni syndrome is associated with germline pathogenic variants in the transcription factor gene *TP53,* and it has been recommended that all ACC cases should be tested for germline pathogenic *TP53* variants (Herrmann *et al.*, 2012).

As previously discussed, germline pathogenic *TP53* variants associated with childhood ACC are particularly prominent in a south Brazilian population (Wasserman *et al.*, 2015). The main pathogenic *TP53* variant hotspots are p.Arg175His, p.Gly245Ser, p.Arg248Gln, p.Arg248Trp, p.Arg273His, and p.Arg282Trp (Wasserman *et al.*, 2015) which affect the DNA binding domain and loops opposing the protein-DNA interacting surface coded for in exons 4-8 (Petitjean *et al.*, 2007). The Brazilian p.Arg337His allele in exon 10 is not one of these common germline pathogenic *TP53* variants, however in a study of 36 paediatric ACC cases in 34 families from the state of Paraná, the pathogenic variant was seen in 97.2% of cases (35 out of 36) (Ribeiro *et al.*, 2001). This high frequency of germline pathogenic *TP53* variants in childhood ACCs has been replicated in other studies, with one predicting an 80% probability that any case diagnosed with ACC before the age of 18 carries a pathogenic *TP53* variant, regardless of family history (Gonzalez *et al.*, 2009). This led to questions about the prevalence of germline pathogenic *TP53* variants in adult ACC cases. Herrmann et al analysed the prevalence of pathogenic *TP53* variants in cases that were diagnosed above 18 years of age (Herrmann *et al.*, 2012). This study identified germline pathogenic *TP53* variants in 3.9% of cases (4 cases out of 103, with age of diagnosis ranging from 21 to 71 years of age), affecting the DNA binding domain (a previously described hotspot region) and tetramerization domain of exon 10 (Herrmann *et al.*, 2012).These studies suggest that pathogenic *TP53* variants are highly penetrant and more likely to have an oncogenic affect early in life.

In contrast to this, Varley et al studied a number of families with childhood adrenocortical tumours and identified inherited *TP53* protein coding variants in unaffected family members (Varley *et al.*, 1999). In addition to this, some variants were identified in family members with late onset cancer. One family identified carried a germline protein coding *TP53* variant in the proband with ACC diagnosed before the age of two years and multiple sarcomas, and in 3rd and 4th degree relatives with breast and stomach cancer, both diagnosed above the age of 40 years (Varley *et al.*, 1999). This study suggests that some low penetrance alleles in *TP53* do exist in ACC cases, and that these might explain some ACC cases without a family history of typical Li Fraumeni syndrome cancers. Additionally, other factors could be contributing to the risk of developing childhood ACC in some *TP53* protein coding variant carriers.

The association between germline pathogenic variants in *TP53* and ACC is by far the most well studied. However, in recent years, germline and somatic pathogenic variants in Lynch syndrome genes have been identified in ACC families. Lynch syndrome is typically associated with an increased risk of colorectal

cancer and is characterised by pathogenic germline variants in DNA mismatch repair genes *MLH1*, *MSH2*, *MSH6*, and *PMS2*. In one study, germline pathogenic variants in Lynch syndrome genes were identified in 3.2% of tested ACC families (3 out of 94 probands), with IHC showing a loss of mismatch repair gene expression in the extracted tumours from these families (Raymond, Everett, *et al.*, 2013). This percentage of ACC affected individuals identified with these germline Lynch syndrome variants is comparable to the prevalence of these variants in colorectal and endometrial cancers, which is estimated to be around 2% and 5% respectively (Raymond, Everett, *et al.*, 2013). The impact of Lynch syndrome in ACC was corroborated by the pan-cancer study of TCGA data, identifying two pathogenic *MSH6* variants and one pathogenic *MSH2* variant in germline DNA from individuals with ACC (Zheng *et al.*, 2016). Additionally, a study of the multiple endocrine neoplasia gene *MEN1* has identified some ACC cases with pathogenic germline variants (Waldmann *et al.*, 2009). However somatic variants in this gene are more common in adrenal tumours, with loss of heterozygosity and copy number variant events being described (Skogseid *et al.*, 1995; Heppner *et al.*, 1999; Zheng *et al.*, 2016).

5.3.7 Aims

This piece of work describes the whole exome sequencing and exploration of variants in germline DNA from individuals diagnosed with ACC. It also explores the gene expression patterns linked to the development of the oncocytic subtype of ACC from TCGA adrenal tumour RNA-sequencing data. The aims are as follows:

1. To identify candidate genes containing germline variants that may affect predisposition to ACC in patients who are negative for pathogenic variants in the known cancer-predisposing gene *TP53*.
2. To identify gene expression changes in RNA-sequencing data from oncocytic ACC tumours that may drive the oncocytic phenotype in comparison to usual type ACCs.

5.4 Materials and Methods

5.4.1 Study Population

Eight individuals diagnosed with ACC were recruited as part of ethically approved studies (REC 12/EE/0478 and REC 14/EE/1059). Five individuals were affected with oncocytic ACC, the remaining three were diagnosed with usual type ACC.

The set is gender balanced (four males and four females); three of the five oncocytic ACC germline samples are from males and the remaining two are from females. The average age at diagnosis was 45 years and was younger for females (35 years of age) than for males (55 years of age). Table 5.1 shows the clinical details for the individuals studied in this analysis.

5.4.2 Germline whole exome sequencing and variant filtering

For all individuals, DNA was extracted from blood and prepared for PE125 WES using the Nextera Rapid Capture Exome enrichment kit (Illumina). Sequencing was performed on HiSeq-4000 machines. The set was called and filtered as part of the in-house dataset and VCF files were generated using a standard germline pipeline following GATK best practice recommendations for WES data (see Chapter 2: Methods for further details).

Common variants that appeared in greater than 5% of European 1000 genomes population were removed from analysis. Protein-affecting variants (loss of function, inframe indels and predicted deleterious and probably damaging missenses (as flagged by SIFT and PolyPhen respectively)) were selected. A total of 1,178 different variants in 1,041 genes appeared in the whole set. An oncocytic specific set was also created for further analysis which included 954 of the above filtered variants in 855 genes.

5.4.3 Variant prioritisation and candidate selection

Variants that passed filtering were manually examined and prioritised to select candidates, removing those that appear in five or more samples from an in-house control set. A literature search was performed for the variant containing gene to identify any known link between that gene and cancer predisposition or the phenotype of interest. The Human Protein Atlas (HPA) (www.proteinatlas.org) (Uhlen *et al.*, 2017) which compiles expression data from GTEx, FANTOM5, and HPA datasets, was used to determine the expression of candidate genes in adrenal tissue. Where data are available for a candidate gene, knockout or allele trapped mouse models were explored for a cancer or adrenal related phenotype using the Mouse Genome Database (www.informatics.jax.org)(Blake *et al.*, 2017). Known cancer predisposing genes including Lynch syndrome genes and *TP53* were searched for protein-affecting, rare variants and any identified were selected as candidates.

Sample ID	Age at diagnosis	Sex	ACC subtype	Mismatch repair Immunohistochemistry
ACC1	67	M	Oncocytic ACC	MSH2/6 preserved
ACC2	33	M	Oncocytic ACC	abnormal pattern
ACC5	55	F	Oncocytic ACC	abnormal pattern
ACC6	45	F	Oncocytic ACC	MSH2/6 preserved
ACC7	63	M	Oncocytic ACC	Not tested
ACC8	21	F	Usual type ACC	Not tested
ACC9	58	M	Usual type ACC	Not tested
ACC10	18	F	Usual type ACC	Not tested

Table 5.1: Details of the oncocytic and usual type ACC samples that underwent WES in this study including immunohistochemistry analysis for mismatch repair deficiencies in extracted tumours where available.

5.4.4 Gene interaction network analysis

To perform computationally intensive gene interaction network analysis, more stringent filters were applied to variants to speed up processing time. The variant allele frequency threshold in 1000 genomes European controls was reduced from 5% to select only variants with an allele frequency of less than or equal to 1%. Variants that are rare in 1000 genomes control set but appear with more than four alternative alleles within in-house control data were removed. Additionally, the top 1% most variable genes within in-house data were removed. This was determined by the number of rare protein-affecting variants each gene contains within the set. The set of 348 different variants were aggregated into 339 genes and input into the Cytoscape GeneMania plugin (Montojo *et al.*, 2010).

Interaction networks were drawn and genes that interacted physically according to GeneMania were clustered. Clusters of four or more genes were further explored. The GO Consortium enrichment analysis web tool was used to apply GO terms to clusters using the PANTHER Overrepresentation Test (release 20171205) including the default false discovery rate (FDR) correction for multiple testing (Blake *et al.*, 2015). Of the terms highlighted by the analysis, the most significant term that encompasses between ten and 200 genes was selected, in compliance with previous studies (Milne *et al.*, 2017).

Allelic counts of all 348 different filtered protein-affecting variants (regardless of GeneMania clustering) within the selected GO terms were aggregated and contingency tables were drawn. Variants were also aggregated for each GO term over a comparably filtered set of 503 Europeans from phase-3 of the 1000 genomes study (Auton *et al.*, 2015). Variants from European phase-3 1000 genomes data were filtered to select 25,021 different rare (European AF <0·01 in 1000 genomes), protein-affecting variants (loss of function, predicted deleterious and damaging missense, and inframe indels). Variants were aggregated into 11,360 genes, which were filtered to remove the top 1% most variable genes. Variability was measured by the number of rare protein-affecting variants each gene contains. Aggregated allele counts were generated for each selected GO term. In total a set of 11,422 genes with variants in 1000 genomes Europeans and the ACC set were analysed for an enrichment of variants in each of the identified GO terms. A one-tailed Fisher's exact test was performed using the R Stats package to test for an enrichment of protein-affecting variants within each selected GO term in ACC in comparison to the European 1000 genomes set.

5.4.5 *TP53* gene interaction analysis

The 339 genes were analysed to identify any that may be closely related to *TP53*. The Cytoscape GeneMania plugin was used to identify genes that cluster with *TP53* via physical interactions. The cluster included genes that were first degree interactors with *TP53* (predicted to have a direct physical interaction) and those that interacted with *TP53* through partner genes, including 'result' genes that were introduced by the Cytoscape GeneMania plugin due to close interactions with the input set. All clustered genes had between one and five degrees of separation (five interaction 'edges') between the candidate

gene and *TP53*. Each gene within the *TP53* cluster was further examined to prioritise candidate variants as above.

A set of 11,360 comparably filtered genes with rare, protein-affecting variants in the 1000 genomes European set (503 individuals) was also analysed for interactions with *TP53*. Of genes input into Cytoscape GeneMania from both the ACC set (339 genes) and the 1000 genomes European set (11,360 genes), the numbers of first and second-degree interactors with *TP53* were counted. A one-tailed Fisher's exact test was used to test for an enrichment of 1st degree interactors with *TP53*, and 1st or 2nd degree interactors with *TP53* in the ACC set. The 2nd degree interactors included those that interacted via a 'result' partner gene which did not have a variant within the input gene set. No 'result' partner genes were counted as first or second-degree interactors.

5.4.6 Tumour immunohistochemistry

The Ventana Benchmark mismatch repair panel (MSH2 (G219-1129) and CONFIRM anti-MSH6) was used to analyse known mismatch repair genes within tumours. IHC was performed on 4µm sections of paraffin embedded tumour tissue in accordance with the manufacturer's guidelines and interpreted by an experienced pathologist (Dr Alison Marker, Department of Histopathology, Addenbrookes Hospital).

5.4.7 The Cancer Genome Atlas – Adrenocortical Carcinoma

The results here are in whole or part based on data generated by TCGA, managed by the National Cancer Institute and the National Human Genome Research Institute. Controlled access data were requested and downloaded for the TCGA-ACC dataset. Aligned germline WES data generated from the blood or tissue derived normal cells of 91 ACC cases as described by Zheng et al (Zheng *et al.*, 2016) were downloaded from TCGA database. FASTQs were generated from BAM files and realigned to hg19 reference genome. Data were run through an in-house VCF generation pipeline and filtered to select rare (1000 genomes European AF < 0.01) protein-affecting variants (loss of function, predicted deleterious and damaging missense variants, and inframe indels). This TCGA set was explored to provide further evidence of the role of candidate genes in this study in ACC predisposition.

To identify any unique gene expression patterns in the oncocytic subset of ACCs, RNA-sequencing data from TCGA were downloaded and run through a purpose-built RNA-sequencing analysis pipeline (see Chapter 2: Methods for further details). The samples were aligned with TopHat (v2.1.1) prior to processing and QC. Differential expression analysis was performed comparing expression levels in the three TCGA oncocytic samples (cases) for which RNA-sequencing data were available, to 74 usual type ACC samples (controls). For each group of cases and controls, an FPKM (fragments per kilobase of transcript per million mapped reads) is calculated by creating a dispersion model which tests if the variance in the group is beyond what is expected from a Poisson distribution model. The model used to calculate the group FPKM value is identical to that used by DESeq (Simon Anders and Wolfgang Huber, 2010). Genes with a significantly different expression value, with an FDR corrected p value (q value) of

less than 0.05, and with a difference in FPKM between the two groups of greater than 1 were further explored. Due to the small number of available cases for this analysis, each sample's FPKM value was manually examined for candidate genes to ensure that the group FPKM is representative of the set.

To understand the functional processes of candidate genes, the GO Consortium enrichment analysis web tool was used to apply GO terms to the set using the PANTHER Overrepresentation Test (version 13.0) web tool. Terms were examined to identify those that might be relevant to the oncocytic subtype.

5.5 Results

5.5.1 Germline whole exome sequencing and variant filtering

After sequencing, a VCF was generated containing 66,447 different variants in the eight ACC individuals. Of these, 60,373 different variants were in the five samples with the oncocytic subtype of ACC. After genotype filtering to remove low quality or potentially misaligned variants (see Chapter 2: Methods for further details), 64,747 different variants were retained in all ACC samples, with 58,922 being retained in the oncocytic subset. A set of 1,178 different rare, protein-affecting variants were retained after variant filtering, of which 954 were in oncocytic samples. These variants were further filtered and aggregated into genes for further analysis.

5.5.2 Gene interaction network analysis

The filtered set of 339 genes input into the Cytoscape GeneMania plugin were aggregated into seven clusters of three or more genes linked by a physical interaction of gene products. Cluster 1 was the largest and consisted of 22 genes, 11 of which were directly interacting with *CUL9* (figure 5.1). The genes were run through GO enrichment analysis to identify a term that could be applied to the set. Cluster 1 was not significantly enriched for any GO term after FDR correction, however the most significantly enriched term with between ten and 200 genes was 'regulation of gene silencing by miRNA' (GO:0060964) ($p = 0.00337$) which was applied to the set. This term contained 81 genes overall, of which only two were seen in cluster one (*NUP153* and *NCOR2).*

Cluster 2 was a smaller interacting set consisting of four genes; *CXCR1*, *BRAT1*, *TTC1*, and *SEC13*. The set also contained the genes *GNAI2* and *ACY1* which were not candidates but were added as interacting partners to the cluster by the Cytoscape GeneMania plugin to allow for the identification of genes that are 2nd degree interactors. No GO terms were identified that contained between ten and 200 genes overall and more than one of the cluster genes. The closest term to matching these criteria was 'organelle localization' (GO:0051640) which contained 577 genes overall, including cluster 2 genes *SEC13* and *BRAT1*. However, this was not applied to the set due to the large overall number of genes included in the term.

Cluster 3 contained eleven genes which centred around an interaction with the ERBB2 interacting protein encoded by *ERBIN*. No significant GO term was identified after multiple testing correction using FDR, however the most significant term prior to correction was 'cell junction assembly' (GO:0034329) ($p = 0.0000476$) which contained 139 genes in total including *PATJ, OCLN*, and *ITGB4* which were identified in cluster 3.

Smaller clusters of three genes were also identified and included within this study. Cluster 4 consisted of genes *GRM1*, *GRK2*, and *MAPK8IP3* with the addition of *MAP3K1* as an interacting partner gene (containing no variants within the ACC set). The most significantly enriched term, containing two of the three genes in cluster 4, was 'activation of MAPK activity' (GO:0000187) ($p = 0.000141$). This

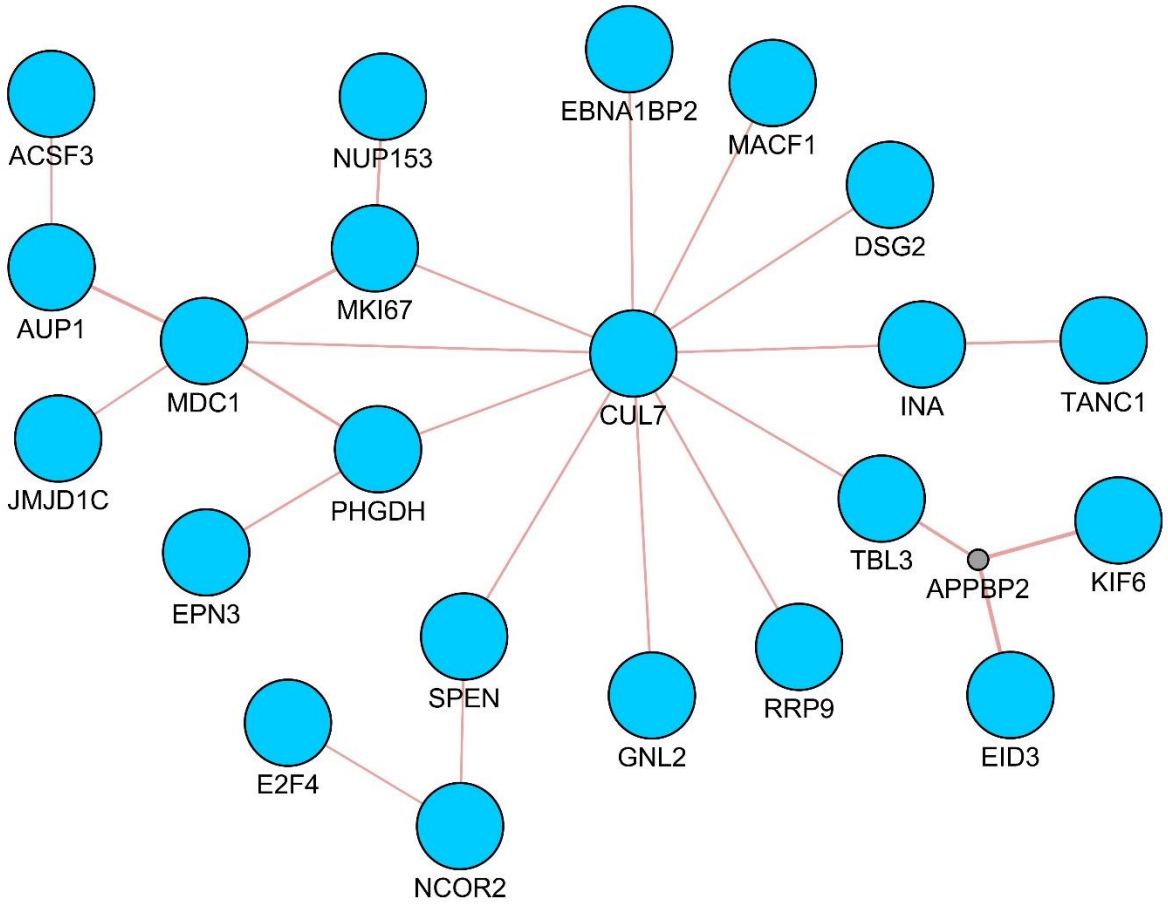


Figure 5.1: Cluster 1 from the gene interaction network analysis showing genes containing rare protein-affecting variants in the ACC set. Interacting partner genes that were not identified with variants in the ACC set are in grey.

term consists of 143 genes including *GRM1*, and *MAPK8IP3*. The three genes in cluster 5 (*SGCB, SGCZ* and *RANGRF)*, were enriched within the term 'muscle cell development' (GO:0055001) (p = 0.000135), with two of the 140 genes in this set being the sarcoglycan genes *SGCZ* and *SGCB.* Cluster 6 (*MEOX2, TDO2,* and *ASMTL*) and cluster 7 (*ERC2, DGUOK,* and *MAPT*) were not associated with any GO terms with between ten and 200 genes.

The four identified GO terms (GO:0060964, GO:0034329, GO:0000187, and GO:0055001) were further explored for an enrichment of rare, protein-affecting variants across the whole ACC set in comparison to a 1000 genomes European control set. Of the 81 genes in the 'regulation of gene silencing by miRNA' (GO:0060964) term, 38 (46.9%) contained variants in either 1000 genomes Europeans or the ACC cases. Across the genes in the term, two variants (aggregated allele count of 2) were identified in ACC and 145 were seen in the control set. This term was not significantly enriched for variants in ACC in comparison to controls (p = 0.695). The term 'cell junction assembly' (GO:0034329) contained variants in 91 of the 139 genes described under the term (65.5%) in the combined case-control set. Five variants were identified in these genes in ACC cases in comparison to the 419 in controls. The set was not statistically enriched for variants within cases (p = 0.864). In the combined set, variants were identified in 80 of the 143 genes (55.9%) in the 'activation of MAPK activity' (GO:0000187) ontology term. This term was not significantly enriched in cases with a combined allele count of 3 in comparison to 280 in controls (p = 0.864). The final tested term, 'muscle cell development' (GO:0055001), was also not significantly enriched in the ACC cases (p = 0.839). The set contained 86 of the 140 genes described under this term and had a combined allele count of 4 in cases and 338 in controls.

None of the identified terms were enriched for variants in ACC cases in comparison to 1000 genome European control samples. However due to the known link between the MAPK pathway and oncogenesis, variants within the cluster termed 'activation of MAPK activity' (GO:0000187) were further explored. Within this cluster, gene *GRM1* contained a rare, protein-affecting missense variant (NM_001278065.1:c.1643C>T, p.Thr548Met, rs201399008). Although this gene contains only one loss of function variant in the ExAC non-TCGA set, it does contain 453 missense variants. A splice donor variant in *MAPK8IP3* within this term was excluded upon manual examination of BAM files due to a lack of reads supporting the variant call, implying that the variant was a sequencing artefact. The G protein-coupled receptor gene *GRK2* (also called *ADRBK1*) was also identified within the cluster however was not in the MAPK related ontology term.

5.5.3 TP53 gene interaction analysis

Of the 339 input genes, ~12% formed an interaction cluster with *TP53*. A set of 50 different variants were identified in 42 genes whose products were predicted to directly or indirectly physically interact with p53. The cluster of *TP53* related genes (figure 5.2) included those with up to five degrees of separation from *TP53* and those that interacted via partner genes. These partner genes are introduced by the Cytoscape

GeneMania plugin to bridge the gap between two secondary interacting genes and are displayed in grey in figure 5.2. None of the identified partner genes were included in the 339 genes used in this analysis, however they allowed other genes to become part of the *TP53* cluster. For example, the protein product of the synaptic membrane exocytosis regulator *RIMS3* interacted physically with the product of the partner gene *BANP*, which directly interacts with *TP53*. Other partner genes introduced to the cluster that act as a direct link between candidate genes and *TP53* included *TAF9* which interacts with *GALNT6*, *COX17* which interacts with *KATNAL1*, and *MAP3K1* which links *GRK2* and *MAPK8IP3* to *TP53*.

The majority of genes in the cluster (29 out of 42, 69%) were first or second-degree interactors with *TP53*, meaning they interacted directly or via another gene. A further seven genes were third degree interactors, with the remaining six genes being fourth, fifth or sixth degree interactors (4, 1, and 1 gene respectively). The number of first and second-degree interactors were counted in comparison to a comparably filtered 1000 genomes European control set. Of the set of 339 input genes with variants in ACC, 11 (~3.2%) were first-degree interactors with *TP53*, 29 (~8.6%) were first or second-degree interactors. The 503 European individuals from phase 3 1000 genomes carried 25,021 different filtered variants in 11,360 genes. Of these input genes, 277 (~2.4%) were first-degree interactors, 3,053 genes (~26.9%) were first or second-degree interactors. There was no statistical enrichment of first or second-degree interactors to *TP53* in the ACC set (p value for 1st interactors: 0.22, 1st and 2nd interactors: 1.00).

Although an enrichment of *TP53* interactors was not seen in the ACC set, individual rare protein-affecting variants within *TP53* interacting genes were further explored to identify any that may confer a risk of ACC development. Within the identified genes, 50 different variants were identified and explored to prioritise candidates that were rare in an in-house control set and were not seen in ExAC or 1000 genomes control sets; 15 different prioritised variants were manually checked in BAM files, of which 11 could be validated. These 11 variants were further explored for a potential role in ACC predisposition. A summary of the 11 different variants identified in *TP53* interacting genes, including counts of missenses and loss of function variants within these genes within the ExAC non-TCGA non-Finnish European (taken from the ExAC non-TCGA control data browser; see Chapter 2: Methods for further details) can be seen in table 5.2.

A germline, loss of function, splice donor variant (NM_000179.2:c.3438+1G>A, rs267608096) in the Lynch syndrome gene *MSH6* was identified in one affected individual. The female (sample ID: ACC5) was diagnosed with oncocytic ACC at the age of 55 and IHC studies on available tumour showed a loss of mismatch repair genes *MSH2* and *MSH6*. This variant has been previously described on ClinVar (Landrum *et al.*, 2018) by the International Society for Gastrointestinal Hereditary

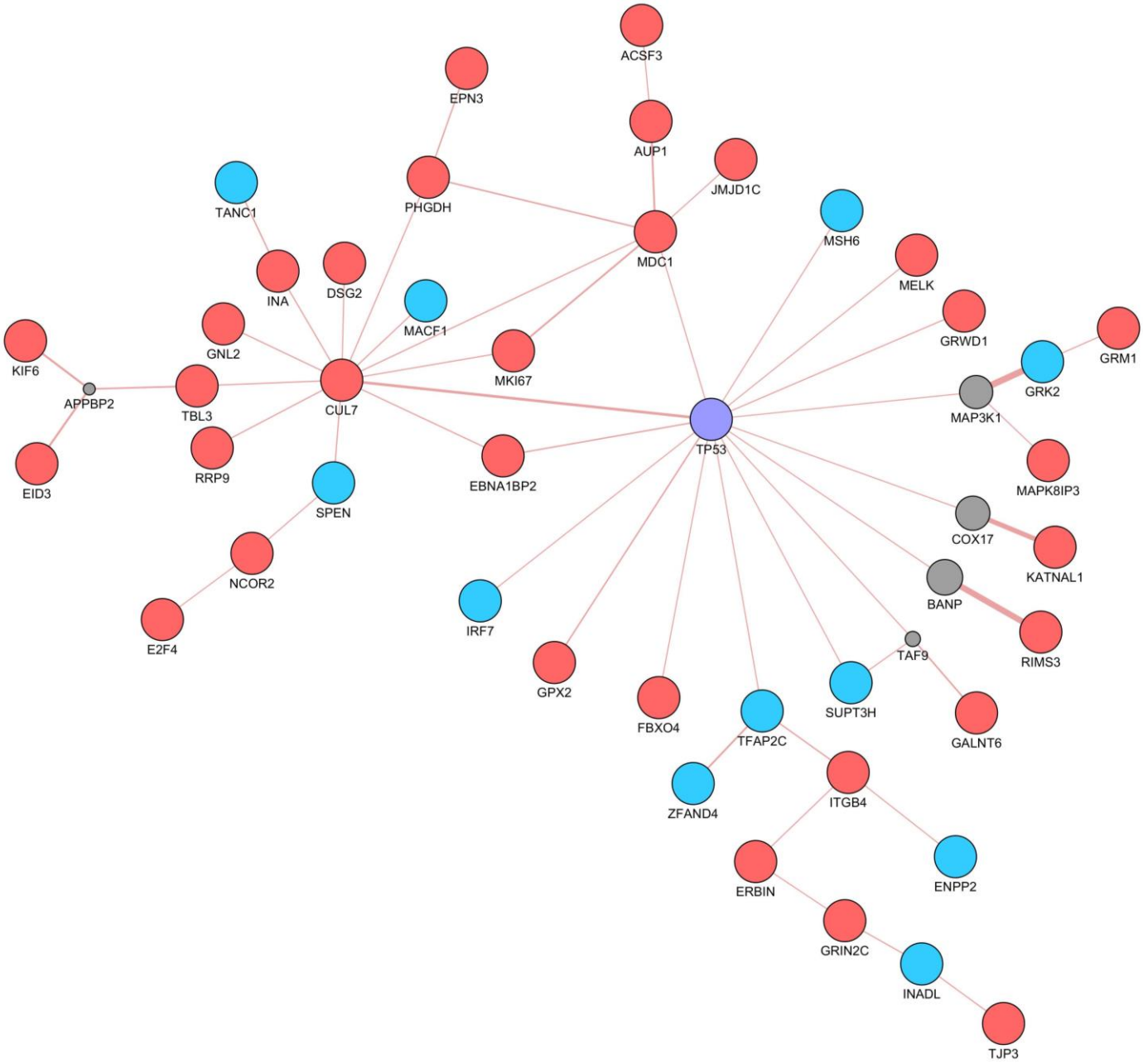


Figure 5.2: A cluster of genes containing rare, protein-affecting variants in ACC individuals that interact directly or indirectly with *TP53* (shown in purple). The 11 candidate genes that were rare in an in-house dataset and appeared real in BAM files are shown in blue. Interacting partner genes that were not identified with variants in the ACC set are in grey.

Gene name	Variant consequence	ID	Amino acid change	Sample	AC LoF variants in this gene in non-TCGA NFE ExAC	AC Missense variants in this gene in non-TCGA NFE ExAC
GRK2	Missense variant	rs201237885	p.Phe85Cys	ACC6	3	33
ENPP2	Missense variant	c.1644C>A	p.Lys548Asn	ACC8	5	632
INADL	Stop gained	c.3964G>T	p.Gly1284*	ACC1	80	16761
IRF7	Missense variant	c.83C>T	p.Gly28Glu	ACC5	35	100
MACF1	Missense variant	rs376859455	p.Phe3171Val	ACC6	7	1505
MSH6	Splice donor variant	rs267608096		ACC5	25	332
SPEN	Missense variant	rs749550885	p.Thr1166Met	ACC2	4	606
SUPT3H	Inframe deletion	rs767570619	p.Cys32_Tyr33del	ACC1	64	46
TANC1	Missense variant	rs143668886	p.Arg1313Gln	ACC1	18	264
TFAP2C	Missense variant	c.1112C>T	p.Thr371Ile	ACC1	0	26
ZFAND4	Frameshift variant	rs773964968	p.Ser355Cysfs	ACC7	101	207

Table 5.2: Candidate variants identified in the ACC set within genes that physically interact with *TP53* according to Cytoscape GeneMania plugin and the frequency of loss of function variants in those genes in ExAC non-TCGA non-Finnish European control set.

Tumours and has been labelled as a 'likely pathogenic' Lynch syndrome variant (ClinVar submission accession: SCV000108066.2).

Other genes containing variants were examined to assess whether the type of variant carried by ACCs is frequent in a healthy population. The previously identified (see 5.5.2 Gene interaction network analysis) MAPK associated G-protein coupled receptor gene *GRK2* (*ADRBK1)* was also identified as associating with *TP53* via *MAP3K1*. The gene carried a rare missense variant (NM_001619.3:c.254T>G, p.Phe85Cys, rs201237885) in a female (sample ID: ACC6) who was diagnosed with an oncocytic subtype ACC at the age of 45. *GRK2* does not frequently contain missense variants in the ExAC non-TCGA non-Finnish European control population, with only 33 protein-affecting missense variants in the set of 27,173 individuals. This low number of variants could be due to the smaller than average gene length (NM_001619.3, 3466 bp, 689aa) or could suggest that variants are associated with disease phenotypes and therefore not present in healthy control populations. This is in contrast to the *TP53* interacting gene *ENPP2*, which contains a novel missense variant (NM_006209.4:c.1644C>A, p.Lys548Asn) in germline DNA from a female with young-onset ACC (sample ID: ACC8), diagnosed at the age of 21. This pyrophosphatase gene is of a similar length to *GRK2* (NM_006209.4, 3233bp, 884aa) however it has a far larger number of protein-affecting missense variants in healthy controls, with 632 missense variants identified in the ExAC non-TCGA non-Finnish European population.

Apart from the aforementioned *MSH6* splice variant, only two other loss of function variants appear in the identified *TP53* interacting genes. A stop gain variant in tight junction associated gene *INADL* (also known as *PATJ*) (NM_176877.3:c.3964G>T, p.Gly1284*) was identified in a male (sample ID: ACC1) diagnosed with an oncocytic ACC at the age of 67. IHC revealed that this individual had preserved MSH2 and MSH6 in the tumour tissue. The ExAC non-TCGA non-Finnish European set contained 80 loss of function variants in this gene; all of these had an AF<0.01. A frameshift deletion of 2bp in zinc finger gene *ZFAND4* (NM_001128324.2: c.1064_1065del, p.Ser355Cysfs, rs773964968), was identified in a male individual (sample ID: ACC7) diagnosed with oncocytic subtype ACC at age 63. In the ExAC non-TCGA non-Finnish European population this gene carries 101 rare (AF<0.01) loss of function variants. The high prevalence of truncating variants seen within these genes in a healthy population suggests that they are not likely to be associated with a disease phenotype.

Variants in *IRF7*, *MACF1*, *SPEN*, *SUPT3H*, and *TANC1* were also excluded from further analysis due to a large number of similar consequence variants seen in a control population. Although genes such as *TANC1* did have a low number of loss of function variants in the healthy population, the variant identified in this study was a missense variant of which there were 264 occurrences in the healthy population. To fully understand the function of these missense variants in both a healthy and disease affected population, they would need to be assessed in the context of the proteins functional domains. A missense variant in transcription factor gene *TFAP2C* was identified in sample ACC1. No loss of function variants and only

26 missense variants have been identified within this gene in the ExAC non-TCGA non-Finnish European population, suggesting this gene is well conserved and could play a role in disease pathogenesis.

5.5.4 Analysis of candidate variants using publicly available datasets and tools

The identified *MSH6* splice donor variant has a known role in Lynch syndrome and can be linked to mismatch repair loss in the tumour of this individual. However, for other identified candidate genes, a link between genotype and phenotype is not as transparent. Freely available online tools can be used to elucidate this further in some cases.

The *TP53* interacting G-protein coupled receptor kinase candidate gene *GRK2* (also known as *ADRBK1*) has not previously been associated with cancer predisposition. Protein and RNA expression data from HPA suggests that this gene is most highly expressed in the bone marrow and immune system including the appendix, tonsil, and spleen. The protein is expressed at a low level in the adrenal gland and GTEx RNA-sequencing data (The GTEx Consortium *et al.*, 2015) summarised by HPA shows an average of 22.6 RPKM (reads per kilobase million) across 145 adrenal samples. This expression level is relatively low, with the highest expression being an average of 153.4 RPKM across 104 spleen samples.

A database of mouse model gene knockouts was used to assess potential phenotypes caused by variants in this gene. A study describing Cre recombination knockout mice showed that deletion of *GRK2*, resulted in developmental retardation and cardiac abnormalities (Matkovich *et al.*, 2006). The variant identified in *GRK2* in ACC was a missense variant and so the phenotype is likely to be less severe than knockout associated phenotypes.

Protein from the transcription factor gene *TFAP2C* was not detected in a number of tissues, including adrenal glands according to HPA data. RNA-sequencing data from GTEx also reflects this lack of expression with an average of 0.2 RPKM across the 145 adrenal samples. This is in comparison to an average RPKM of 17 in oesophageal tissue and 29.6 in skin tissue within GTEx data. Mouse models in this gene have highlighted a role in early development, specifically proliferation and differentiation of trophectodermal cells (Werling and Schorle, 2002).

5.5.5 Candidate genes in The Cancer Genome Atlas – Adrenocortical Carcinoma dataset

The candidate genes *GRK2*, *MSH6*, and *TFAP2C* were further explored in the TCGA-ACC dataset. No variants were found in this set in candidate gene *GRK2*, however one missense was identified in *GRK3* as well as a missense in *GRK1*, *GRK4*, *GRK5*, *GRK6*, and two missenses in *GRK7*.

A frameshift insertion was seen in two unrelated individuals from the TCGA-ACC set in the mismatch repair gene *MSH6* (NM_000179.2:c.4065-4066T>TTTGA, p.Lys1358Aspfs, rs267608142). Other mismatch repair variants were seen in this set in *MSH2* (NM_000251.2:c.4G>A, p.Ala2Thr, rs63750466 and NM_000251.2:c.2156T>G, p.Leu719Trp, rs777933557) and *MSH3* (NM_002439.4:c.2732T>G, p.Leu911Trp, rs41545019). No variants were seen in candidate gene *TFAP2C*.

5.5.6 Differential expression analysis in oncocytic The Cancer Genome Atlas data

RNA-sequencing data from three oncocytic samples within the TCGA subset were analysed in comparison to data from 74 usual type ACC samples to identify genes or sets of genes (defined by ontology terms) that were differentially expressed. Within this analysis, usual type ACC samples were used as controls to identify oncocytic specific expression changes. No RNA sequencing data derived from normal tissue was available from these samples to act as an additional 'true' control. Expression levels across the group were tested for 63,652 genes and RNAs. Of these, 127 were significantly differentially expressed between the two sets (FDR corrected p value < 0.05). A set of 61 genes and RNAs had an FPKM change of greater than 1.

These genes were assigned GO terms using the GO Consortium enrichment analysis web tool. A set of 42 ontology terms were highlighted that contained less than 200 genes in total, with each term carrying more than one of the 61 candidate genes. Out of the 42 highlighted terms the candidate gene *OXT*, which codes for the hormone Oxytocin, was in 22 terms and was significantly differentially expressed between the tested sets (p = 0.00005). This gene was included in, but was not limited to the terms 'renal system process' (GO:0003014), 'response to fatty acid' (GO:0070542), 'regulation of blood pressure' (GO:0008217), and 'regulation of synapse organization' (GO:0050807). The *OXT* gene had expression levels of 0 FPKM within the oncocytic subset and 1.31 FPKM in usual type ACCs, the significant difference between the values is driven by the 0 FPKM. Of the 74 usual type ACCs included as controls, 44 carried an FPKM of 0 for the *OXT* gene (figure 5.3A) and therefore it is unlikely to be associated with the oncocytic subtype. Similarly, the CART prepropeptide coding gene, *CARTPT*, is significantly differentially expressed in cases and controls (p = 0.00005), with an FPKM of 0 in the oncocytic cases and 2.09 in the usual type ACC controls. This gene was identified in 17 of the highlighted GO terms but also has an FPKM of 0 in 38 controls.

One interesting term highlighted through this analysis was 'negative regulation of Wnt signaling pathway' (GO:0030178) which could have implications in oncogenesis. The two genes identified within this term with a significant expression change were *SFRP5* (p = 0.00005) and *NFATC4* (p = 0.0003). Of these two genes, nuclear factor gene *NFATC4* had the largest expression change, with an FPKM of 16.67 in cases and 118.60 in controls. Two oncocytic ACC tumour samples had extremely low expression of this gene, with individual FPKMs of 0.99 and 2.80, the final case had an FPKM of 46.21 which was more comparable to the values seen in controls. Gene expression values for *NFATC4* in the usual type ACC controls ranged from 0.75 FPKM, lower than the minimum FPKM in cases, to 577.39 FPKM, with a mean of 118.60 and a standard deviation of 128.20. In spite of this high deviation from the mean, the data distribution does vary between the oncocytic ACC subtype and usual ACC controls (figure 5.3B). In comparison, *SFRP5*, had a smaller change in expression with an FPKM of 0 in cases and 2.35 in controls. Within this gene, 33 controls also had an FPKM of 0, however one sample did carry an FPKM of 156.31,

likely contributing to the high group average. The distribution of FPKM in cases and controls for *SFRP5* can be seen in figure 5.3C.

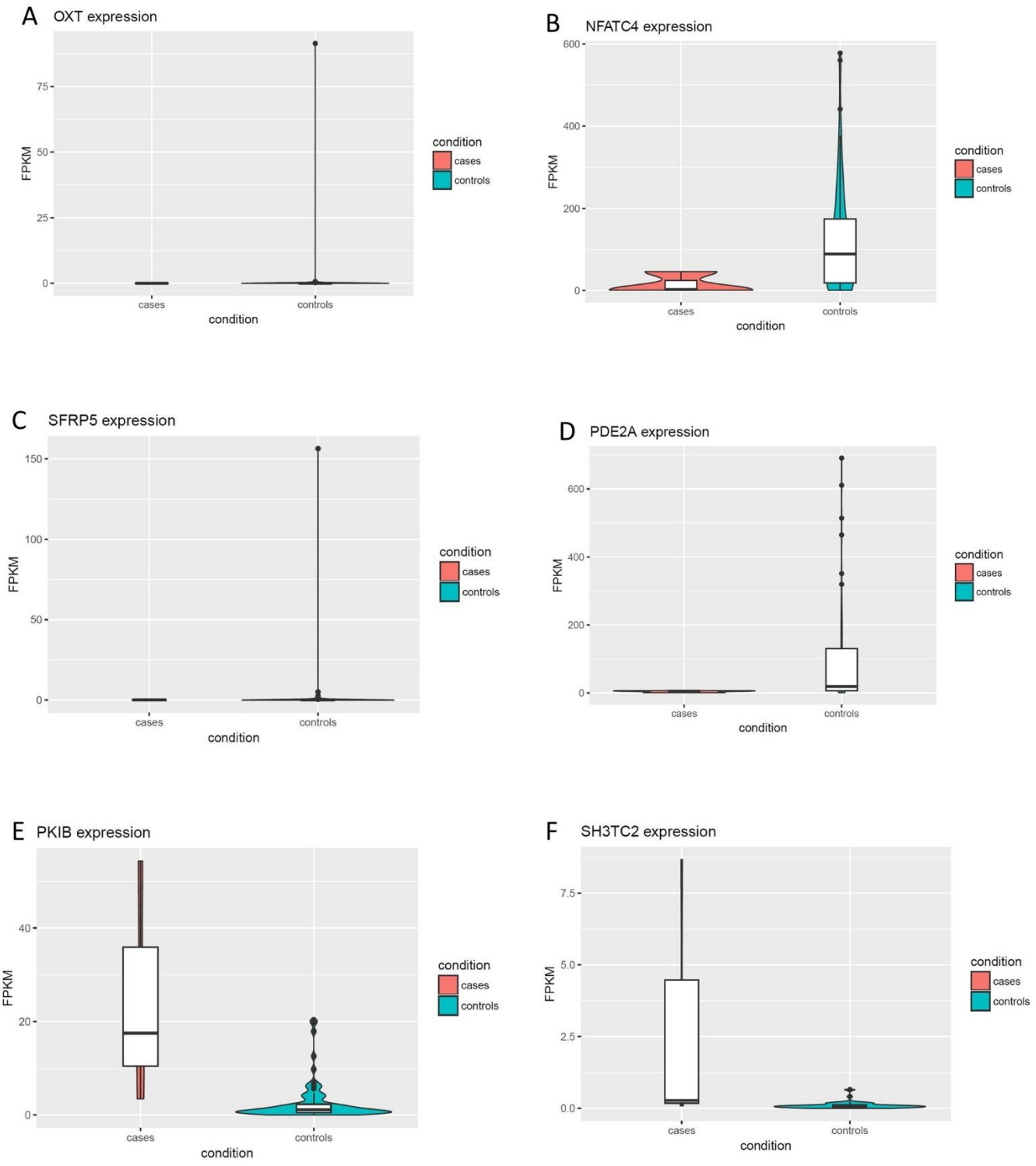


Figure 5.3 A-F: Candidate genes with a differential gene expression in oncocytic ACC samples (cases) in comparison to usual type ACC samples (controls). Gene expression is shown in fragments per kilobase of transcript per million mapped reads (FPKM).

Another highlighted gene was *PDE2A* which showed a significant loss of expression in the oncocytic ACC subtype cases (p = 0.00005) with an FPKM of 4.40 in cases and 91.06 in controls. In the oncocytic ACC cases, expression ranged from 0.94 to 6.93 FPKM; in comparison, expression in the controls ranged from 1.12 to 690.37 FPKM (figure 5.3D). This gene was identified in six GO terms; 'response to purine-containing compound' (GO:0014074), 'response to cAMP' (GO:0051591), 'carbohydrate derivative catabolic process' (GO:1901136), 'response to organophosphorus' (GO:0046683), 'organophosphate catabolic process' (GO:0046434), and 'purine-containing compound catabolic process' (GO:0072523).

Of the 61 significantly differentially expressed genes analysed, 57 (~93%) showed a loss of expression in the oncocytic ACC subtype case set. Of the loss of expression genes, 48 (~84%) had a value of 0 FPKM in the case set. Of the four identified gain of expression genes, two had expression levels of 0 FPKM in the usual type ACC control set. The genes *AL901608.1* and *IGHV1OR16*-4 had no expression in any of the usual type ACC controls, nor in two of the three cases. Therefore only 11 genes had expression values in both cases and controls and are likely to represent true expression changes. These included the two candidate loss of expression genes *NFAT4* and *PDE2A.* The two gain of expression genes in this set included *PKIB* (p = 0.00015), with 25.10 FPKM in cases and 2.68 FPKM in controls (figure 5.3E), and *SH3TC2* (p = 0.0002) with 3.01 FPKM in cases and 0.10 FPKM in controls (figure 5.3F).

5.6 Discussion

Germline WES data from individuals with usual type or oncocytic ACC were analysed to identify any predisposing genetic factors. GO analysis was used to highlight clusters of functionally related genes containing rare germline protein-affecting variants that could be affecting disease predisposition. Four GO terms were identified and further explored for an enrichment in ACC cases. The first Cytoscape GeneMania generated cluster consisted of 22 genes, however the GO term assigned to this, 'regulation of gene silencing by miRNA', was only associated with variants in *NUP153* and *NCOR2*. This GO term did not contain the other 20 genes shown to physically interact in the gene cluster, showing a discrepancy in genes that share a GO term and those that are interacting physically according to Cytoscape GeneMania. None of the identified terms in this analysis, which also included 'cell junction assembly', 'activation of MAPK activity', and 'muscle cell development', were statistically enriched for rare protein-affecting variants in ACC cases in comparison to 1000 genome European controls.

The MAPK signalling pathway is often altered in tumours and has been shown to be downregulated in adrenal tumours (Rubin *et al.*, 2015). Genes that were identified with variants in the 'activation of MAPK activity' set were further explored, revealing a rare protein-affecting missense variant in the conserved gene (as determined by its variability in control sets) *GRK2* which was later also identified as interacting with *TP53*. This G protein-coupled receptor gene has been linked to cancer development and prognosis in a number of different settings. The *GRK2* gene has been described as a tumour suppressor in thyroid cancer, with overexpression showing a reduction in cell proliferation in cell lines (Métayé *et al.*, 2008). In contrast to this, an overexpression of this gene has been seen in a number of tumour environments including breast, thyroid, pancreatic, and prostate tumours, causing to be linked to an increase in proliferation and poor overall survival (Prowatke *et al.*, 2007; Métayé *et al.*, 2008; Nogués *et al.*, 2016; Zhou *et al.*, 2016). Although *GRK2* has currently only been shown to be somatically altered in cancers, the low level of germline variability of this gene in healthy controls suggest it is well conserved, with only three loss of function and 33 missense variants identified in ExAC non-TCGA non-Finnish Europeans (combined allele count of 36 out of 27,173 individuals) and one missense variant seen in the 1958 birth cohort.

GRK2 has low levels of expression in a healthy adrenal mouse gland and mice knockouts were shown to cause developmental and cardiac abnormalities (Matkovich *et al.*, 2006). Studies of *GRK2* and the related gene *GRK5* showed that although both products interact with p53, only GRK5 can phosphorylate and regulate the tumour suppressor (Wei *et al.*, 2012). In contrast to this, a separate study revealed a dose-dependent correlation between *GRK2* expression and levels of phosphorylated p53, causing a delay to cell cycle progression in hepatocellular carcinoma cells (Wei *et al.*, 2012). Interestingly, missense variants in other G protein coupled receptor genes including *GRK1, GRK3, GRK4, GRK5, GRK6,* and *GRK7* were identified in the ACC set analysed from TCGA data including usual and oncocytic types.

Pathogenic germline variants in Lynch syndrome genes have been previously identified in 3.2% of ACC cases (Raymond, Everett, *et al.*, 2013). The mismatch repair gene *MSH6*, the product of which was found to directly interact with p53, contained a rare loss of function variant in germline DNA from one individual with oncocytic ACC. Tumour material from this individual was also shown to have loss of *MSH2* and *MSH6* via IHC. Although loss of *MSH2* and *MSH6* have been shown to be caused by germline defects in *MSH6* in colorectal tumours (Morak *et al.*, 2017), within this study tumour material from a second individual revealed loss of *MSH2* and *MSH6* without the presence of any germline protein-affecting variants in mismatch repair genes.

Of the *TP53* interacting genes highlighted within this study, only two others (apart from *MSH6*) contained loss of function variants. The variants identified in the genes *INADL* and *ZFAND4* were rare (1000 genomes European AF < 0.01), however overall the genes carried a high number of loss of function variants in the ExAC non-TCGA control population. Genes are described as 'tolerant' to loss of function variants if they harbour a higher than expected number of these variants, suggesting that they are not evolutionarily conserved nor under a high degree of negative selective pressure. Genes may be under negative selective pressure if changes to that gene impact key cellular pathways and cause an increased risk of disease. Therefore, genes with a high number of loss of function variants in a healthy population are less likely to cause an increased risk of disease predisposition.

A missense variant in the transcription factor gene *TFAP2C* was identified in ACC1. Although this gene does not show high levels of variability in the ExAC non-TCGA non-Finnish European controls, suggesting a high level of conservation, it also has very low to no expression in healthy adrenal samples (mean RPKM of 0.2 across 145 healthy adrenal samples in comparison to a mean RPKM of 29.6 in healthy skin samples) and so a missense variant is unlikely to have a deleterious effect in this tissue. No further variants were identified in this gene in the TCGA-ACC subset.

Analysis of expression data from ACC samples within TCGA revealed a potential role for *NFATC4*, *PDE2A*, *PKIB*, and *SH3TC2* in oncocytic ACC development. The protein product of the gene *NFAT4C* is involved in the Wnt signalling pathway. Wnt-pathway activation had been previously shown in TCGA-ACC data in comparison to Wnt wild-type samples, in addition to somatic variants in the Wnt/β-catenin pathway in 41% of tested ACC cases (Zheng *et al.*, 2016). The nuclear factor of activated T cells (NFAT) family of transcription factors are activated by the Wnt signalling pathway in a calcium dependent manner, causing the expression of downstream Wnt target genes (Van Camp *et al.*, 2014). In correlation with findings by Zheng et al (Zheng *et al.*, 2016), findings here suggest that there is an increased activation of the Wnt signalling pathway in ACC tumour samples, as determined by the high expression of *NFATC4* with an FPKM of 118.60. Interestingly, this high expression is not seen in oncocytic ACC tumour samples, with expression in this subset much lower at 16.19 FPKM (comparable to values seen in healthy

adrenal samples in GTEx data), suggesting that unlike the usual type ACC, Wnt signalling is not hyperactivated during oncocytic ACC oncogenesis.

Cyclic adenosine monophosphate (cAMP) is a component of the cAMP/PKA signalling pathway that regulates cellular differentiation, growth, cell metabolism, and cell death. The phosphodiesterase encoded by *PDE2A* is activated by cyclic guanosine monophosphate (cGMP) and is involved in the degradation of both cAMP and cGMP (Stroop and Beavo, 1991). By decreasing phosphodiesterase activity, and therefore decreasing degradation of cAMP, one can produce a mild hypercorticosteronemia phenotype in mice (de Joussineau *et al.*, 2012). Phosphodiesterases are described as modifiers of abnormal adrenal phenotypes instead of being directly causal of adrenal tumours. For example, loss of heterozygosity of *PDE11A4* has been linked to the development of adrenal hyperplasia in adrenocortical tumour patients (Horvath *et al.*, 2006). Additionally, germline exonic variants in *PDE11A* are a modifying factor for the development of testicular and adrenal tumours in individuals carrying germline variants in *PRKAR1A* (Libé *et al.*, 2011). A reduction of phosphodiesterase expression could therefore cause aberrant cAMP pathway signalling and in some cases lead to cAMP-regulatory element mediated transcriptional activity causing excessive cortisol production (Libé and Bertherat, 2005; de Joussineau *et al.*, 2012).

Neither of the two candidate genes *PKIB* or *SH3TC2* were associated with a GO term. There is little known about the gene *SH3TC2* besides its involvement in Charcot-Marie-Tooth disease (Brewer *et al.*, 2014). In contrast, *PKIB* is a well-studied protein kinase inhibitor. These genes were highlighted as the only two genes to show a gain of gene expression in the oncocytic ACC subset after filtering. The product of gene *PKIB* is also linked to cAMP signalling as a member of the cAMP-dependent protein kinase inhibitor family. The cAMP-protein kinase A (PKA) pathway is involved in the control of cell growth and differentiation in almost all mammalian tissues (Taskén and Aandahl, 2004). *PKIB* has previously been shown to be overexpressed specifically in aggressive and castration resistant prostate cancer (Chung *et al.*, 2009). *PKIB* is involved in PKA-C nuclear localisation and modification of kinase activity causing Akt phosphorylation and activation (Chung *et al.*, 2009).

The secretion of cortisol is a key diagnostic factor of adrenal tumours. The cAMP pathway is a regulatory process of cortisol secretion, with cAMP-PKA activation stimulating cortisol production and release. The two genes with a change in expression linked to cAMP appear to collectively and individually promote the cAMP-PKA pathway (Zheng *et al.*, 2016), contrary to this, clinical information for these three individuals show that the extracted oncocytic adrenal tumours were not hormone secreting. This could suggest that cAMP-PKA is contributing to different oncogenic processes being utilised in the oncocytic subtype of ACC in comparison to the usual subtype.

5.7 Summary

This study contains one of the largest described in-depth analyses of the genetic aspects of oncocytic and usual type ACC. In-house generated WES data was utilised to describe a number of germline candidate

variants in these samples, although no clearly associated novel candidates were identified. The majority of participants in this sequencing study were diagnosed over the age of 40 and none were described with a family history of ACC, therefore the probability of finding causal germline mutations in this set and comparable sporadic cancer datasets is low and the results are perhaps unsurprising. In addition to using in-house WES data to identify potential candidates for ACC predisposition, external data were utilised to explore possible pathways of oncocytic subtype development. G-protein coupled receptor genes were identified with germline variants within in-house and TCGA ACC data, suggesting a potential role for the MAPK pathway in disease predisposition. Analysis of RNA-sequencing data suggested that the cAMP-PKA pathway may have a specific role in the oncogenesis of the oncocytic subtype and therefore may, be a potential drug target.

6 Predisposition to hereditary breast cancer

6.1 Introductory statement

The work described in this chapter is the joint effort of myself and other researchers within the group. All WES preparation was performed by James Redman and Alexey Larionov and sequenced at the Cancer Research UK Cambridge Institute Genomics Core. Data were processed to VCF by an in-house WES pipeline written by Alexey Larionov. Sequencing data from the Montreal dataset were processed and analysed by me. The WECARE WES data were processed to VCF by Alexey Larionov. Variant filtering for the WECARE set was performed separately by both Alexey Larionov and myself. The GO enrichment analysis was designed by myself and performed on the Montreal set and on my own comparably filtered version of the WECARE set. Statistical tests on the WECARE set were recommended by David Conti and performed by Alexey Larionov using his own comparably filtered variant set. Further prioritisation, including breast cancer interaction analysis, on a subset of WECARE candidate genes was performed by myself.

6.2 Abstract

Breast cancer is the most common cancer diagnosed in the UK. Hereditary breast cancers are often associated with germline variants in *BRCA1* and *BRCA2*. Although other genes have been identified containing breast cancer predisposing variants, these only account for a small minority of the remaining *BRCA1/BRCA2*-negative hereditary breast cancer cases. Individuals with contralateral breast cancer have been shown to be enriched for hereditary predisposition in comparison to unilateral breast cancer cases. Additionally, ethnically homogenous populations such as the Ashkenazi Jews have been shown to be at higher risk of hereditary diseases, largely due to the presence of a number of disease associated founder variants. This study explores whole exome sequencing data from the Women's Environmental Cancer and Radiation Epidemiology (WECARE) study, looking at contralateral breast cancer as a model for increased breast cancer risk, and from a set of familial and Ashkenazi Jewish breast cancer cases recruited via the McGill Cancer Genetics Programme in Montreal. Both datasets were explored for an enrichment of loss of function variants in GO terms and individual variants within enriched terms were further explored using a manual variant prioritisation approach. Further statistical tests were performed on the WECARE set to prioritise candidate genes, which were explored for an interaction with known breast cancer predisposing genes. Within the Ashkenazi Jewish families from the Montreal set, two different rare, loss of function variants were identified in DNA recombination related gene *RAD52*. Within the whole Montreal familial breast cancer set, four GO terms were significantly enriched for loss of function variants in comparison to a 1000 genomes control set and an in-house control set including; 'cysteine-type endopeptidase activity' (GO:0004197), 'single-stranded RNA binding' (GO:0003727), 'magnesium ion binding' (GO:0000287), and 'cholesterol binding' (GO:0015485). After statistical prioritisation of the WECARE set, variants were identified in genes that interacted with known breast cancer genes. In particular, variants within the gene encoding STK11

interacting protein *(STK11IP)* were seen in both the WECARE and Ashkenazi Jewish individuals in the Montreal set. A number of candidate genes have been suggested that, when carrying pathogenic variants, could explain some of the missing heritability seen in familial breast cancer cases.

6.3 Introduction

6.3.1 Hereditary breast cancer syndromes

Breast cancer is the second most common cause of cancer mortality in women in the UK (Cancer Research UK, 2016). Between 5-15% of breast cancers are thought to be hereditary (Carroll *et al.*, 2008; Economopoulou, Dimitriadis and Psyrri, 2015). Around 30% of hereditary breast cancers are associated with germline pathogenic variants in genes *BRCA1 and BRCA2* (Siegel, Naishadham and Jemal, 2013). A prospective cohort study reports a wide range of risk estimates for carriers of pathogenic variants in these genes, with a cumulative risk by the age of 70 of between 40% to 87% and 27% to 84% for *BRCA1* and *BRCA2* respectively (Kuchenbaecker *et al.*, 2017). It has been shown that individuals with a family history of breast cancer can carry at least a twofold risk of disease development (depending on the strength of the family history), with pathogenic variants in *BRCA1* and *BRCA2* accounting for around 15% of this excess risk (Easton, 1999; Antoniou *et al.*, 2008).

In recent years, other breast cancer predisposition genes including *PALB2*, *CHEK2*, and *ATM* have been incorporated into the BOADICEA risk score model which calculates *BRCA1* and *BRCA2* carrier probabilities for individuals with family histories of breast cancer (Lee *et al.*, 2016). Pathogenic variants in these genes are predicted to increase risk of breast cancer (by age of 80) to 50% for *PALB2*, 30% for *CHEK2*, and 28% for *ATM* (Lee *et al.*, 2016). These three genes, like *BRCA1* and *BRCA2* are part of the DNA damage repair pathway.

Pathogenic variants in *TP53* have been described in Li Fraumeni syndrome, which was originally identified as conferring a risk to childhood rhabdomyosarcoma, sarcoma, leukaemia, brain tumours, breast cancer, and adrenocortical carcinoma (Li and Fraumeni, 1969). Germline pathogenic variants in this tumour suppressor gene are associated with a small fraction of breast cancer cases, although breast cancers do account for around one third of cancers in Li Fraumeni syndrome families (Birch *et al.*, 1998).

PTEN Hamartoma Tumour Syndrome (PHTS) is associated with germline variants in *PTEN*. Pathogenic variants in this gene increase risk of endometrial, thyroid, renal, and colorectal cancers in addition to breast cancer. Lifetime breast cancer risk estimates for germline pathogenic *PTEN* variant carriers range from 67-85% (Tan *et al.*, 2012; Bubien *et al.*, 2013; Nieuwenhuis *et al.*, 2014).

Despite the multiplicity of breast cancer predisposing genes, including many that have not been mentioned, families with a strong history of breast cancer who are negative for germline variants in these known cancer genes are often seen. This suggests that there are still more risk associated genes to identify. Studies looking to identify novel breast cancer predisposition genes or variants often utilise ethnically homogenous populations and those who carry a greater cancer risk such as those with multiple primary tumours. The Ashkenazi Jewish population is the largest genetically isolated population in the United States and is often used to identify founder variants that are rarer and less

easily identifiable in the general population (Rinella *et al.*, 2013; Carmi *et al.*, 2014). An example of this is the identification of Crohn's disease risk variants which were found to have up to 4-fold higher prevalence in Ashkenazi Jewish individuals (Kenny *et al.*, 2012). Additionally this technique was successfully utilised by Rinella et al, to identify new pathogenic variants in *BRCA1* and *BRCA2* negative Ashkenazi Jewish breast cancer cases (Rinella *et al.*, 2013).

6.3.2 Contralateral Breast Cancer

The development of multiple primary tumours is thought to be indicative of a greater overall predisposition for cancer (Cybulski, Nazarali and Narod, 2014). Evidence for this can be found by looking at populations of individuals who develop no cancer, one cancer or multiple cancers. This latter group of individuals is larger than would be expected by chance (Cybulski, Nazarali and Narod, 2014), suggesting that there is a greater probability for some individuals to develop a second primary tumour. Additionally, it has often been described that the sites of the second tumour occurrence are specific and in these cases can often be linked to a hereditary syndrome (Lynch *et al.*, 1977; Cybulski, Nazarali and Narod, 2014). This is particularly the case with multiple breast tumours, which is known as contralateral breast cancer (CBC) when the first and second primary tumours appear in different breasts.

CBC can be described as synchronous, when the second tumour is discovered close in time to the first, or asynchronous, when there is a larger (although currently undefined) period of time between the first and second diagnosis (Narod, 2014). These differences are important to distinguish between tumours that develop as a result of the primary tumour metastasising to the other breast, and those that are a true second primary tumour. Synchronous breast cancer is less common, constituting around 30% of multiple breast tumour cases (Hartman *et al.*, 2005); they can be further categorised by whether the symptoms appear in both breasts or in one breast, the latter of which is usually only discovered as a result of staging investigations following the discovery of the first tumour (Narod, 2014). It should be noted that within the latter group, some second primary tumours may have not been identified during this screening and may therefore have been labelled as asynchronous. There is no standard for classifying synchronous or asynchronous BC, however the term synchronous can generally be applied to second primary tumours that appear less than a year post-diagnoses of the first tumour (Senkus *et al.*, 2014). To create a clear distinction between multiple primary tumours and potential metastatic secondary tumours, studies exploring CBC often consider only asynchronous cases, with a clear distinction in time between diagnosis of the first and second primary tumour (Bernstein *et al.*, 2004).

Pathogenic variants in *BRCA1* and *BRCA2* are well described as contributors to hereditary breast cancer as previously mentioned (Kuchenbaecker *et al.*, 2017). Pathogenic variants in both genes also confer an elevated risk for second primary CBC, with risk estimates of 41% for *BRCA1* and 21% for *BRCA2* within 20 years of diagnosis of the first tumour (Kuchenbaecker *et al.*, 2017). However, only 5% of CBC patients carry pathogenic variants in *BRCA1* or *BRCA2* (Hartman *et al.*, 2005; Malone *et al.*, 2010). Studies have examined genetic risk factors for CBC in *BRCA1* and *BRCA2* pathogenic variant

negative individuals with a strong family history (Bernstein *et al.*, 2004; Sisti *et al.*, 2015). The WECARE study identified a modest increased risk to CBC in carriers of a protein-truncating *CHEK2* variant, 1100delC (Mellemkjær *et al.*, 2008). However this variant has been described as rare in most populations and unlikely to account for a large proportion of CBC cases (Mellemkjær *et al.*, 2008).

6.3.3 Tumour Characteristics and CBC risk

When describing the relationship between the characteristics of the first tumour and the risk of developing CBC, it is important to note that different tumour subtypes are subject to different treatments. It can therefore be a challenge to distinguish between the effect of tumour status and the effect of differential treatment. For example, estrogen-receptor (ER) positive tumours have been associated with a decreased risk for CBC development which is thought to be due to the standard treatment of this tumour type with adjuvant endocrine therapy (Lizarraga *et al.*, 2013).

A similar trend was observed in cases with ER negative and PR (progesterone-receptor) negative first primary tumours, who had a higher risk of developing CBC to ER+/PR+ cases (Saltzman *et al.*, 2012). It was also shown in the same study that women with HER2-overexpressing or triple negative first primary tumours were at higher risk of CBC development in comparison to women with a luminal-A subtype (Saltzman *et al.*, 2012). In line with previous suggestions that this difference in risk could be linked to the treatment given, the risk became non-significant when analysis was limited to patients who were not treated with adjuvant hormonal therapy (Saltzman *et al.*, 2012). Another interesting observation with regards to hormone receptor (HR) status is that there is a correlation between the molecular subtype seen in the first and second primary tumours; for example, women whose first primary tumour is HR-negative are around 10 times more at risk of developing a second HR-negative tumour (Kurian *et al.*, 2009). This suggests that similar genetic or environmental factors are contributing to the risk and development of the first and second primary tumours (Saltzman *et al.*, 2012).

Studies have been published looking into the effects of lobular histology, medullary histology and high histologic grade on the development of CBC and have shown to slightly increase risk (Lizarraga *et al.*, 2013). However risk rates with regards to lobular histology of the first primary tumour have been shown to differ amongst studies (Lizarraga *et al.*, 2013; Langlands *et al.*, 2016). This could be a result of differences in methods of diagnosis, interpretation of histological results, and different degrees and methods of surveillance and disease management between study centres as well as general clinical approaches in different countries (Langlands *et al.*, 2016). One study suggested that there is no increased incidence of CBC in patients with invasive lobular carcinoma or invasive ductal carcinoma (Langlands *et al.*, 2016). However when studies were limited to cases of synchronous CBC, there was a higher incidence of lobular cancers than seen in UBC cases (Kollias *et al.*, 2001). This result has similarly been seen in other population studies (Bernstein *et al.*, 1992) and therefore a lobular histology has been described as a risk factor for developing a second synchronous primary breast cancer.

The observation that often one BC subtype is reflected in both the first and second primary tumours gives further credence to the idea that both tumours are likely to be a result of the same combination of risk factors such as predisposition genes. The link between predisposing gene and tumour characteristic is best seen in patients with *BRCA1* pathogenic variants who are at greater risk of being diagnosed with triple-negative breast cancers (Pellegrino *et al.*, 2016).

6.3.4 The effect of first primary cancer treatment on CBC risk

Treatment of the first primary breast cancer has long been thought of as a possible risk or protective factor against the development of subsequent tumours. The widespread adoption of tamoxifen as a therapy for ER-positive patients in 1985 has been linked to a 3% decrease in cases of CBC each year (Lizarraga *et al.*, 2013). In one study the use of hormonal therapy has been linked to a decrease in CBC cases by 42% (Schaapveld *et al.*, 2008). Chemotherapy was also shown to decrease CBC risk by 27% (Schaapveld *et al.*, 2008).

The effects of radiotherapy on the development of a second primary breast cancer have been extensively studied. Radiation in doses comparable to those delivered by radiotherapy has been shown to induce breast cancer, particularly in young women (under 24 years of age) (Boice Jr. *et al.*, 1991). Therefore the dose given to the primary BC has been described as influencing CBC development (Boice Jr. *et al.*, 1992). The increased risk of developing CBC after radiotherapy was established previously but was shown to contribute to less than 3% of CBCs (Boice Jr. *et al.*, 1992). This risk was similarly shown to be increased in younger women whose first BC was treated before the age of 45 (Boice Jr. *et al.*, 1992). The same relationship between CBC and radiotherapy in younger women was shown in the WECARE study (Stovall *et al.*, 2008), which was originally designed with the intention of analysing the interaction between radiotherapy, pathogenic variants in known BC predisposition genes and the development of CBC (Bernstein *et al.*, 2004). Many of the known BC predisposition genes play key roles in DNA damage repair, including the repair of double strand breaks induced by ionizing radiation. It was therefore hypothesised that individuals with pathogenic variants in BC associated DNA damage repair genes may be more susceptible to radiation induced CBC (Bernstein *et al.*, 2004). The WECARE study later showed that radiotherapy for the first BC and specific variants in the double strand break repair gene *ATM* increase the risk of developing second primary breast cancers (Bernstein *et al.*, 2010).

6.3.5 Aims

This study uses high risk breast cancer populations to identify new breast cancer associated predisposition genes. The aims are as follows:

1. To identify candidate variants for an association with a predisposition to CBC development
2. To explore variants in an Ashkenazi Jewish breast cancer cohort to identify potential founder variants with an association with breast cancer risk

6.4 Methods

6.4.1 Study Population

Three populations of samples were used for this analysis. In total, 512 germline DNA samples were utilised for sequencing from the WECARE (Women's Environmental Cancer and Radiation Epidemiology) study (Bernstein et al. 2004). This included 256 unilateral breast cancer (UBC) and 254 CBC patients and two patients with an unknown phenotype that were later removed from analysis. WECARE was created as a case-control study matching 1,398 patients with UBC as controls to 705 CBC patients. Participants were recruited through five cancer registries covering Denmark and the US within Iowa, California (Los Angeles and San Diego) and Washington; all registries from within the US were associated with the Surveillance, Epidemiology and End Results (SEER) registry system (Malone et al. 2010).

Participants were eligible for the WECARE study as cases if they meet the following conditions as described in the study design (Bernstein et al. 2004):

a) The first primary invasive breast cancer did not spread beyond the regional lymph nodes at the time of diagnosis.

b) The second primary breast cancer must be diagnosed at least one year after the diagnosis of the first breast cancer.

c) The participant resided in the same reporting area for both diagnoses.

d) The participant had no previous cancer diagnosis.

e) The participant was under the age of 55 at the time of the first primary breast cancer diagnosis.

f) The participant was alive at the time of contact to provide consent, complete an interview and provide a blood sample.

Participants from WECARE were selected for the study described here if they met the following additional criteria:

a) Participant does not carry any pathogenic variants in *BRCA1*, *BRCA2* or *PALB2*.

b) Participant was of white ethnicity with European ancestry.

c) The germline DNA sample passed quality control in a previous genome wide association study (Bernstein et al. 2004).

d) The germline DNA sample could be matched to a control on the use of radiotherapy on the first primary tumour.

Clinical information about the study participants was made available including time between diagnosis of first and second primary tumour where appropriate and treatment given for first primary tumour. All participants were recruited through registries in the USA or Denmark and are ethnically matched to generate a group of predominantly American women of European ancestry.

As an external control, data from 198 female germline DNA samples from phase 3 of the 1000 genomes project were aggregated into the WECARE set. Ethnicity tests using eigenvectors were used to confirm the ethnicity of the WECARE set and to select non-Finnish European samples from the 1000 genomes project that match this ethnicity. In addition to the WECARE study, 63 women with breast, ovarian or pancreatic cancer were recruited from centres in Montreal via the McGill Cancer Genetics Programme. All recruited individuals had a strong family history of breast cancer. A BRCAPro score, which is based on studies of Ashkenazi Jewish and European descent individuals, was generated to predict the likelihood of families carrying pathogenic variants in *BRCA1* or *BRCA2* (Huo *et al.*, 2009). Individuals with a BRCAPro score of greater than 10%, but with no pathogenic variants in these genes were selected. Of this set, 14 individuals were of Ashkenazi Jewish ancestry, and so in addition to analysis across all 63 cases, this small subset was analysed separately.

6.4.2 Germline whole exome sequencing

DNA was extracted from blood and prepared for PE125 WES using the Nextera Rapid Capture Exome enrichment kit (Illumina). For the WECARE study, sequencing was performed on HiSeq-2500 machines by the CRUK CI genomics core facility. Germline DNA from the Montreal breast cancer participants was sequenced on HiSeq-4000 machines.

VCF files for both sets were generated using a standard pipeline following GATK best practice recommendations for whole exome data (see Chapter 2: Methods for further details). For the WECARE set, FASTQ files from 198 non-Finnish European females from the 1000 genomes project (NFE) were integrated for joint variant calling to reduce the likelihood of technical differences affecting association testing. Genotypes were filtered to select those with a genotype quality score of greater than 20 and a genotype depth of less than 500 to remove incorrectly aligned regions. A variant call rate filter was set separately for WECARE samples with NFEs, where each variant had to be present in greater than or equal to 85% of WECARE genotypes and 85% of NFE genotypes. A WECARE only set was also created with more permissive filters, selecting variants that appeared in greater than or equal to 85% of WECARE genotypes only for analysis where the NFE set was not used.

The WECARE dataset was filtered to select uncommon variants (AF <0.05 or > 0.95 (to allow for rare variants that are found in the reference genome) in either the WECARE set only or the WECARE set including NFEs), protein-affecting variants (loss of function, predicted deleterious and damaging missense (as flagged by SIFT and PolyPhen respectively)) and those that were predicted to be

**The WECARE dataset underwent comparable filtering by bioinformatician Alexey Larionov, prior to statistical gene prioritisation as further described which was performed by Alexey Larionov and David Conti.*

pathogenic by ClinSig (flags included: “likely_pathogenic”, “risk_factor”, “pathogenic", "association", "protective", "drug_response”). Variants that were predicted to be “benign” or “likely_benign” by ClinSig were removed regardless of predicted variant consequence*.

Standard genotype filters were used on the Montreal set, including selecting those with a genotype quality score of greater than 20 and a genotype depth of less than 500 to remove incorrectly aligned regions, in addition to a variant call rate filter of greater than or equal to 50%. Variants were filtered to select uncommon (AF <0·05 in European 1000 genomes), protein-affecting variants (loss of function, predicted deleterious and damaging missense (as flagged by SIFT and PolyPhen respectively), and inframe indels).

6.4.3 Gene ontology enrichment analysis

Similar to previous analysis on the HDGC dataset (see Chapter 3), GO terms were used to identify clusters of genes enriched for variants in each breast cancer cohort in comparison to a 1000 genomes control set. A multi-use R markdown script was developed that takes a dataset of interest, filters according to parameters set by the user, creates an aggregated count of variants in a specified set of GO terms and runs a Fisher’s exact test to compare this count to that of a control set.

The script accepts Rdata, with one dataframe for VEP annotated variants and another for genotype information as is generated by the in-house pipeline. Parameters that need to be set by the user include a maximum allele frequency (AF) filter and a population which ethnically matches the case dataset, both the case and 1000 genomes dataset are then filtered for variants with an allele frequency below the filter in the set population from 1000 genomes. Additionally, users can set a consequence filter of either “LOF”, stating that only loss of function variants should be used, or “PAV”, allowing the analysis to be performed on all protein-affecting variants (loss of function, predicted deleterious and damaging missense variants and inframe indels). The Montreal set was filtered for loss of function variants with an AF < 0.05 in the European 1000 genomes population.

The case and control sets are filtered accordingly, and allele counts for each variant are aggregated into GO terms. A GO annotation file (GAF) containing terms and associated gene IDs was downloaded from the GO data site. Terms that in total contained between 20 and 200 genes and contained variants within these genes in either the case or control set were analysed. A one-tailed Fisher’s exact test was applied to select terms that were significantly enriched for variants in the case set, with an FDR corrected p value of less than 0.01. Variants identified within the enriched terms are output separately for further exploration.

The same script was applied to test for GO terms associated with breast cancer risk in the Ashkenazi Jewish Montreal subset and the WECARE set. All sets were tested for an enrichment of

† Described statistical tests were recommended by statistician David Conti and implemented by Alexey Larionov using variant sets filtered comparably to those previously described.

GO terms consisting of aggregated loss of function variants with an allele frequency of less than 0.05 in the European 1000 genomes cohort. A script was later developed comparing the case set of interest to an in-house sequenced control set of mixed cancer types. This script checks that the case set is not included in the in-house set (as this consists of all samples sequenced within the local sequencing facility) and recalculates allele counts and frequencies of the control set if necessary (excluding variants within case individuals) before performing the same enrichment analysis for GO terms.

Directed acyclic graphs of significantly disease associated GO terms were drawn using the European Bioinformatics Institute web tool QuickGO (Binns *et al.*, 2009).

6.4.4 WECARE gene prioritisation and targeted sequencing†

Three types of statistical analysis were performed on the WECARE set to prioritise genes for further study; 1) a sequence kernel association test (SKAT) with burden based variant aggregation, 2) a SKAT with variance based variant aggregation, 3) a proportional odds logistic regression model. A set of filtered variants were input into burden and variance based SKATs (using R SKAT package), using a generalised linear regression model to test for an association between variants and risk of CBC. For this test, variants were aggregated into 8,649 gene based functional regions using burden SKAT, which assumes that all variants detected in one gene have the same direction of effect on the phenotype and so are all protective or all risk causing, and variance based SKAT which allows variants to have opposite directions of effect. Although burden based SKAT tests harbour more statistical power when all variants aggregated affect the phenotype in the same direction with a similar order of magnitude, this power is dramatically decreased when aggregated variants have opposing effects on phenotype (e.g. some variants increase risk while others are protective) (Lee *et al.*, 2012). In comparison, the variance based method tests and aggregates individual variant effects and weights, allowing for the detection of rare genes which harbour both protective and risk causing variants which would otherwise remain undetected and may be important influencers of CBC predisposition (Lee *et al.*, 2012). Covariates included in this analysis include age at diagnosis, number of pregnancies and the top two calculated eigenvectors.

A proportional odds logistic regression (using R MASS package) was performed to test for an association between aggregated genes (using burden based SKAT) and CBC in comparison to both UBC and unaffected NFE samples. For all tests, genes were ranked by p value. Genes with a p value of less than 0.05 in one of the three tests were selected for further analysis. Genes with a p value of less than 0.05 in two out of the three tests were automatically selected for targeted sequencing in a greater cohort (neither the methods nor results of the additional targeted sequencing will be described within this thesis).

† *Described statistical tests were recommended by statistician David Conti and implemented by Alexey Larionov using variant sets filtered comparably to those previously described.*

6.4.5 Breast cancer predisposition gene interaction analysis

Gene interaction analysis was performed to identify variants within genes that perform a similar function or interact with breast cancer predisposition genes. Twelve breast cancer predisposition genes were selected from a consensus paper by Easton et al (Easton *et al.*, 2015) (*BRCA1, BRCA2, TP53, PTEN, CDH1, STK11, NF1, PALB2, ATM, CHEK2, NBN,* and *RECQL*). These risk genes, in addition to the prioritised WECARE genes, were input into the Cytoscape GeneMania plugin (Montojo *et al.*, 2010). A cluster of genes that either directly or indirectly interacted with the twelve risk genes, either by physical interaction or pathway interaction, was selected.

The filtered variants within the identified directly interacting candidate genes were explored manually to identify any potential link to disease predisposition. In particular, the frequency of variants in the candidate genes were assessed in the ExAC non-TCGA control cohort. This method was based upon the assumption that genes with a low variability in a healthy population are more likely to be disease causing when mutated.

6.5 Results

6.5.1 Germline whole exome sequencing and variant filtering

The VCF generation pipeline produced a set of 343,824 different variants across the 710 germline samples in the WECARE set, including 198 non-Finnish European (NFE) 1000 genomes samples. After genotype filtering as described, variants were filtered on call rate across the WECARE set only and the WECARE set including the NFE samples, retaining 298,610 different variants and 130,135 different variants respectively. Although the WECARE set specifically included individuals without *BRCA1*, *BRCA2* or *PALB2* variants, a number of protein-affecting variants in these genes were identified, and so 11 individuals carrying these variants were removed from analysis, additionally two samples that were identified as ethnic outliers after PCA were removed from the study. The variants were filtered for consequence and rarity in each set, retaining 18,683 different variants that were in WECARE samples and 11,362 different variants in the WECARE and NFE set.

Raw sequencing data of the 63 Montreal individuals contained 131,895 different variants. Genotype and call rate filtering reduced this number to a set of 130,757 different variants which were further filtered on variant consequence and rarity to produce a set of 6,074 different variants.

6.5.2 Gene ontology enrichment analysis

Within the WECARE study, the previously filtered set of 18,683 different variants in the WECARE only set were input into GO enrichment analysis. Variants were aggregated into 9,087 genes, which were compared to filtered 1000 genomes control variants. Within the 1000 genomes control set, 200,693 different protein-affecting variants were filtered down to loss of function variants with a European allele frequency of less than 0.05. After filtering, 20,969 different variants were retained in the control set and aggregated into 9,823 genes.

Variant counts were aggregated into 1,971 GO terms and were tested for an enrichment in cases vs controls using a Fisher's exact test for each term. Tests on the WECARE set revealed no terms that were significantly enriched in cases in comparison to the control set. The most significant FDR corrected p value was 0.13 for the term "protein binding involved in protein folding" (GO:0044183).

Variants within the Montreal breast cancer set were further filtered by the GO script to select 478 different loss of function variants within 447 genes that were aggregated into 1,971 GO terms as before. Of the tested GO terms, 14 were significantly enriched (FDR corrected $p < 0.01$) for loss of function variants in the case set (table 6.1). Descriptors of these terms showed three were associated with cell adhesion, including 'cell-cell junction assembly' (GO:0007043), 'catenin complex' (GO:0016342), 'cell-cell adhesion mediated by cadherin' (GO:0044331).

GO ID	Term description	Namespace	Counts for minor alleles in Montreal cases	Counts for minor alleles in controls	Counts for major alleles in Montreal cases	Counts for major alleles in controls	P value	FDR corrected P value
GO:0004197	cysteine-type endopeptidase activity	molecular function	74	632	0	374	5.34E-15	1.1E-11
GO:0071556	integral component of lumenal side of endoplasmic reticulum membrane	cellular component	44	265	24	741	2.05E-10	0.0000002
GO:0003727	single-stranded RNA binding	molecular function	31	150	37	856	8.45E-09	0.0000056
GO:0032728	positive regulation of interferon-beta production	biological process	28	141	40	865	1.51E-07	0.000074
GO:0000287	magnesium ion binding	molecular function	43	322	25	684	3.37E-07	0.00011
GO:0034080	CENP-A containing nucleosome assembly	biological process	12	22	56	984	2.86E-07	0.00011
GO:0015485	cholesterol binding	molecular function	26	142	42	864	2.24E-06	0.00063
GO:0005882	intermediate filament	cellular component	26	148	42	858	4.6E-06	0.0011
GO:0036126	sperm flagellum	cellular component	14	46	54	960	7.68E-06	0.0017
GO:0007043	cell-cell junction assembly	biological process	7	8	61	998	1.29E-05	0.0023
GO:0016342	catenin complex	cellular component	7	8	61	998	1.29E-05	0.0023
GO:0003725	double-stranded RNA binding	molecular function	20	102	48	904	2.17E-05	0.0036
GO:0004601	peroxidase activity	molecular function	24	144	44	862	2.93E-05	0.0044
GO:0044331	cell-cell adhesion mediated by cadherin	biological process	6	6	62	1000	3.54E-05	0.005

Table 6.1: Gene ontology terms that were enriched for loss of function variants in the Montreal set in comparison to 1000 genomes controls (FDR corrected p value < 0.01)

The enriched GO terms could be split into three main namespaces (classes of GO terms), biological processes, cellular components, and molecular functions although some terms overlap multiple namespaces. Directed acyclic graphs were drawn for terms within the three namespaces. The four enriched biological processes were 'positive regulation of interferon-beta production' (GO:0032728), 'CENP-A containing nucleosome assembly' (GO:0034080), 'cell-cell junction assembly' (GO:0007043), and 'cell-cell adhesion mediated by cadherin' (GO:0044331) (figure 6.1). Enriched cellular component terms included 'catenin complex' (GO:0016342), 'integral component of lumenal side of endoplasmic reticulum membrane' (GO:0071556), 'intermediate filament' (GO:0005882), and 'sperm flagellum' (GO:0036126) (figure 6.2). Six enriched terms were labelled as molecular functions including 'cysteine-type endopeptidase activity' (GO:0004197), 'single-stranded RNA binding' (GO:0003727), 'magnesium ion binding' (GO:0000287), 'cholesterol binding' (GO:0015485), 'double-stranded RNA binding' (GO:0003725), and 'peroxidase activity' (GO:0004601) (figure 6.3).

The variants identified within these enriched terms were explored and it was discovered that two different variants appeared in a number of different Montreal samples and could be driving enrichment in their corresponding ontology terms. The two frameshift insertion variants in *MIS18BP1* (NM_018353.4: c.471_472insT, p.Leu158IlefsTer8, rs546807245) (figure 6.4) and *MNS1* (NM_018365.2: c.605delA, p.Lys202SerfsTer9, rs549395315) (figure 6.5) were checked in an in-house control set and were found to be carried by other germline non-breast cancer samples. Both variants were indels, which occasionally can be sequencing artefacts, often called in repetitive regions. Therefore, BAM files were manually checked for these variants in samples where the variant was called heterozygous, homozygous reference and missing after genotype filtering. With both variants, the presence of insertions and deletions at the base of interest, in addition to the presence of the variant in non-called samples and at repetitive regions suggested that these were sequencing artefacts. However, the *MNS1* variant was present in 43% of reads in the sample in which it was called as heterozygous, in contrast to the other two samples where it was not called which had ~7% and ~6% of reads with a deletion. Although in the case of this sample, this variant may not have been a sequencing artefact, it is possible that variant calling is generating artefacts in this repetitive region. This emphasises the need to manually check BAM files for all candidate variants.

To reduce the enrichment associated with sequencing artefacts, GO enrichment analysis was performed comparing to an in-house control set. Four GO terms were enriched both in comparison to the 1000 genomes control set and the in-house control set. These molecular function terms were 'cysteine-type endopeptidase activity' (GO:0004197), 'single-stranded RNA binding' (GO:0003727), 'magnesium ion binding' (GO:0000287), 'cholesterol binding' (GO:0015485). The 14 different rare, loss of function variants identified in these terms are described in table 6.2.

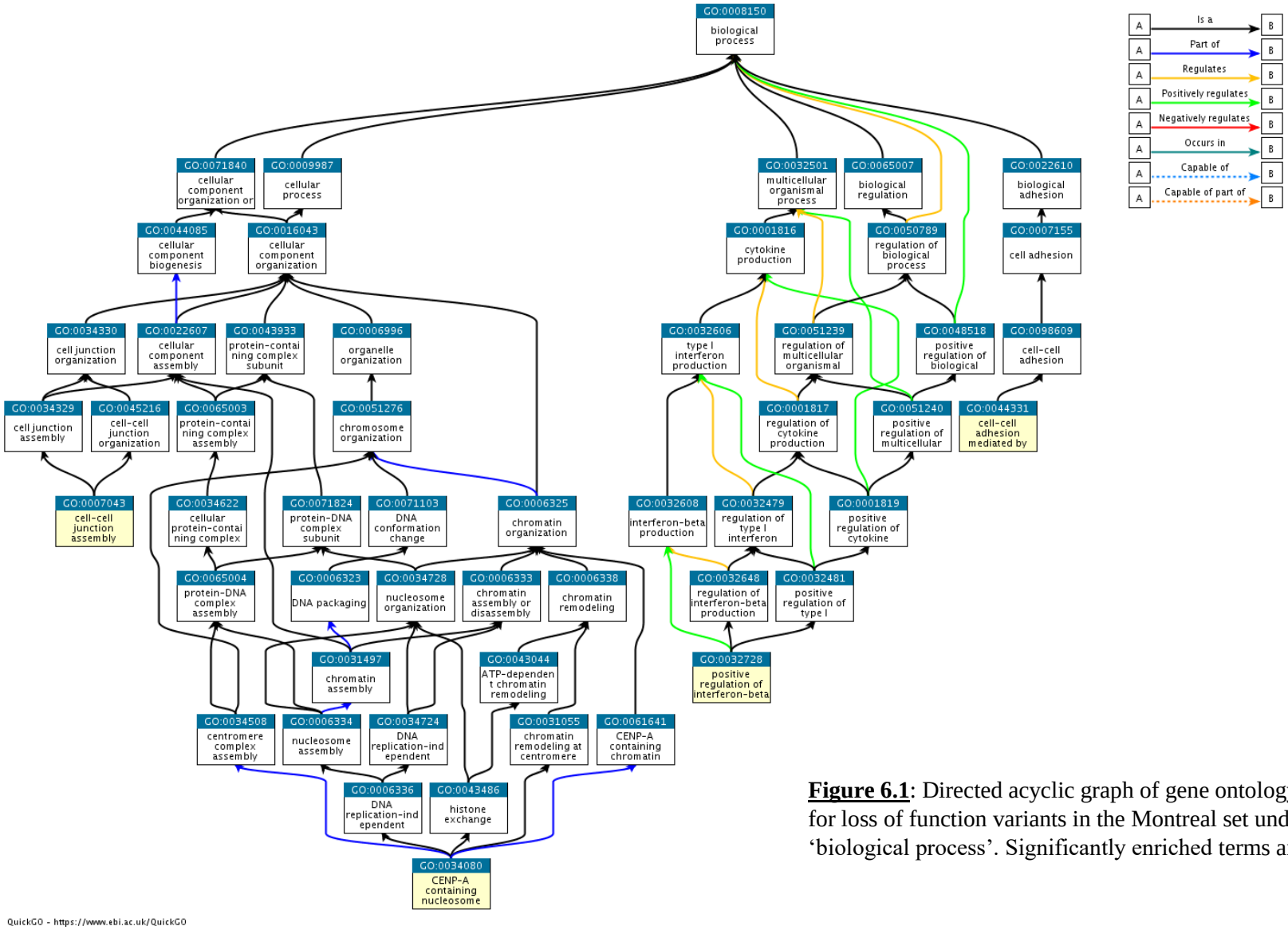


Figure 6.1: Directed acyclic graph of gene ontology terms significantly enriched for loss of function variants in the Montreal set under the namespace of 'biological process'. Significantly enriched terms are shown in cream.

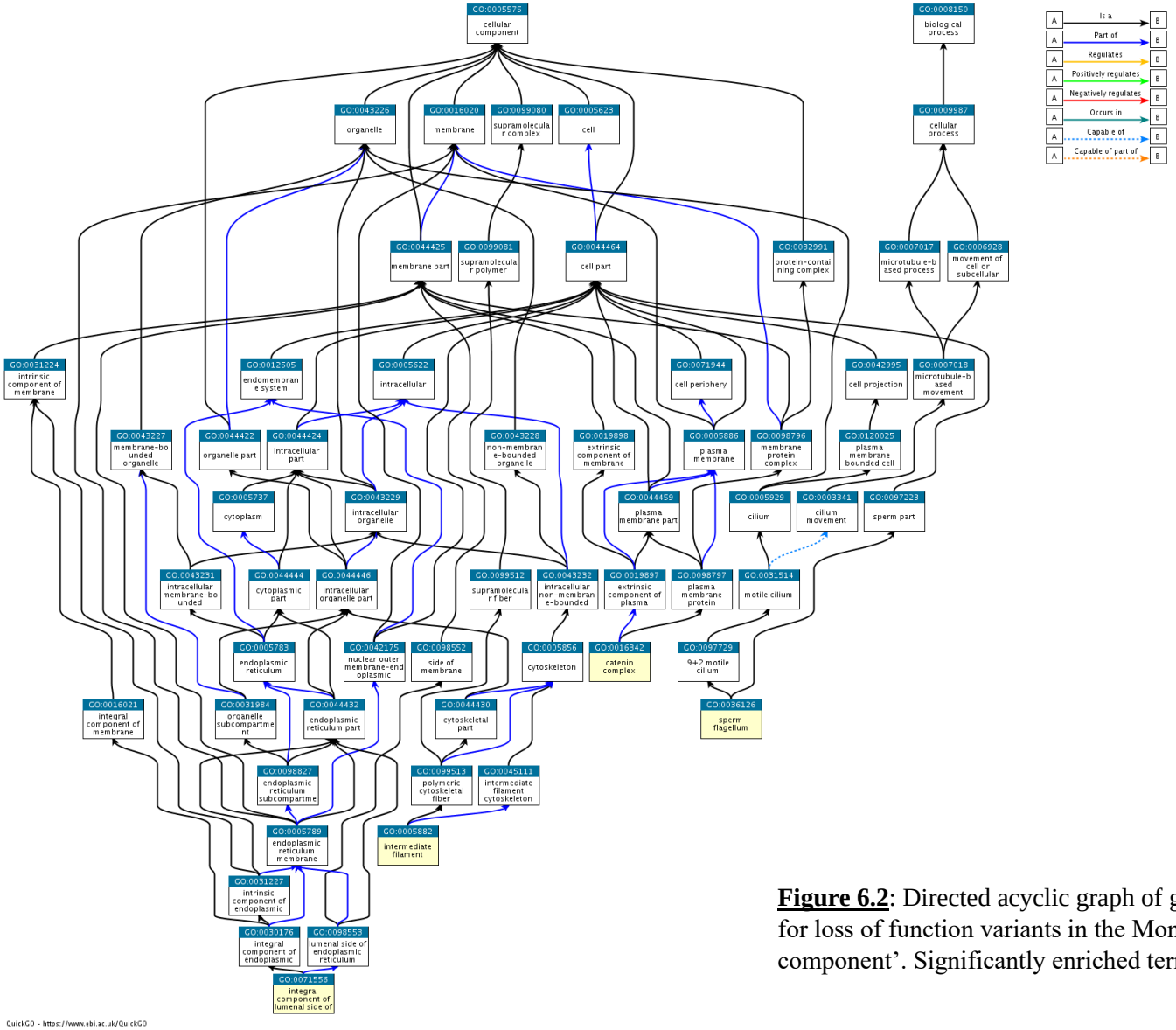


Figure 6.2: Directed acyclic graph of gene ontology terms significantly enriched for loss of function variants in the Montreal set under the namespace of 'cellular component'. Significantly enriched terms are shown in cream.

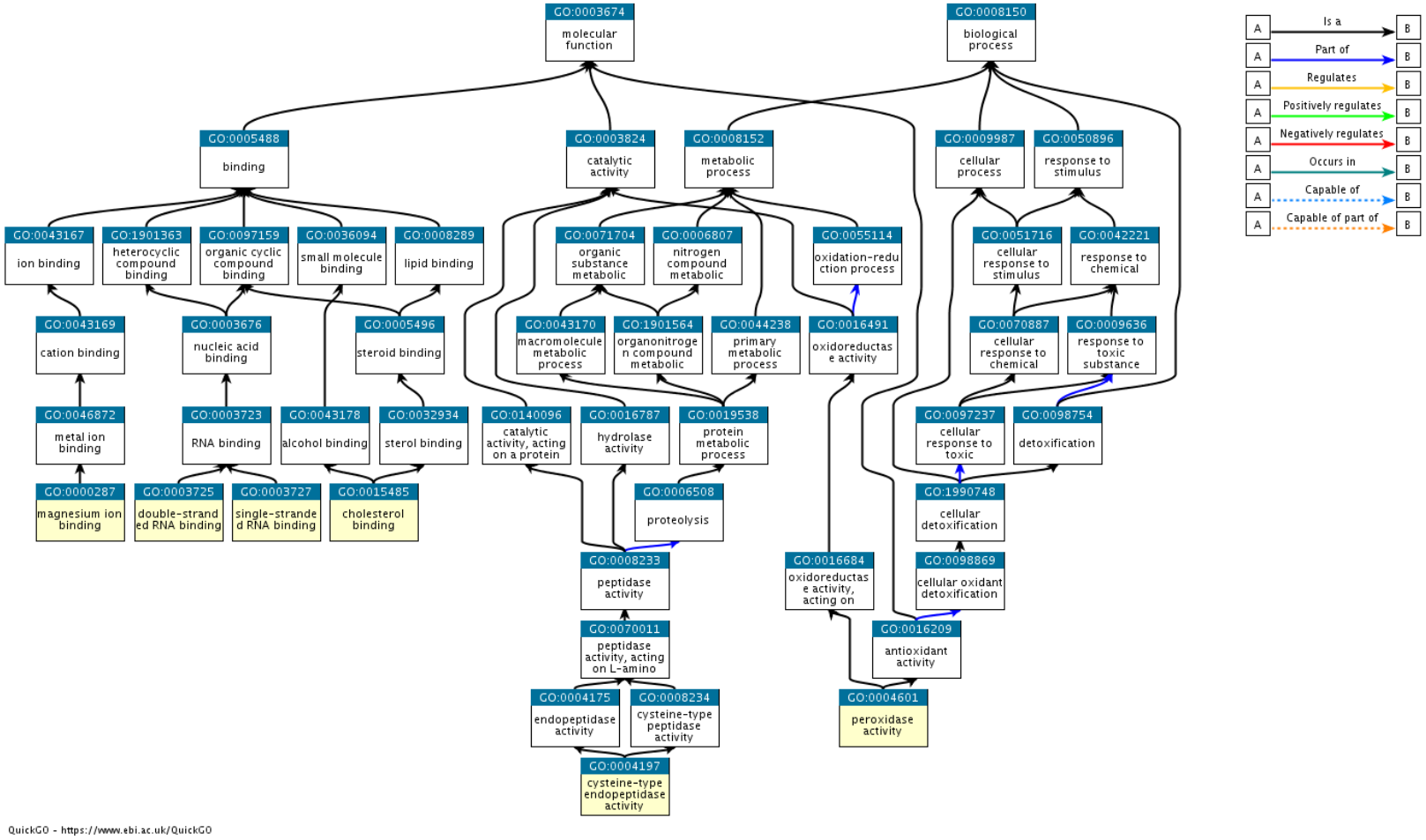


Figure 6.3: Directed acyclic graph of gene ontology terms significantly enriched for loss of function variants in the Montreal set under the namespace of 'molecular function'. Significantly enriched terms are shown in cream.

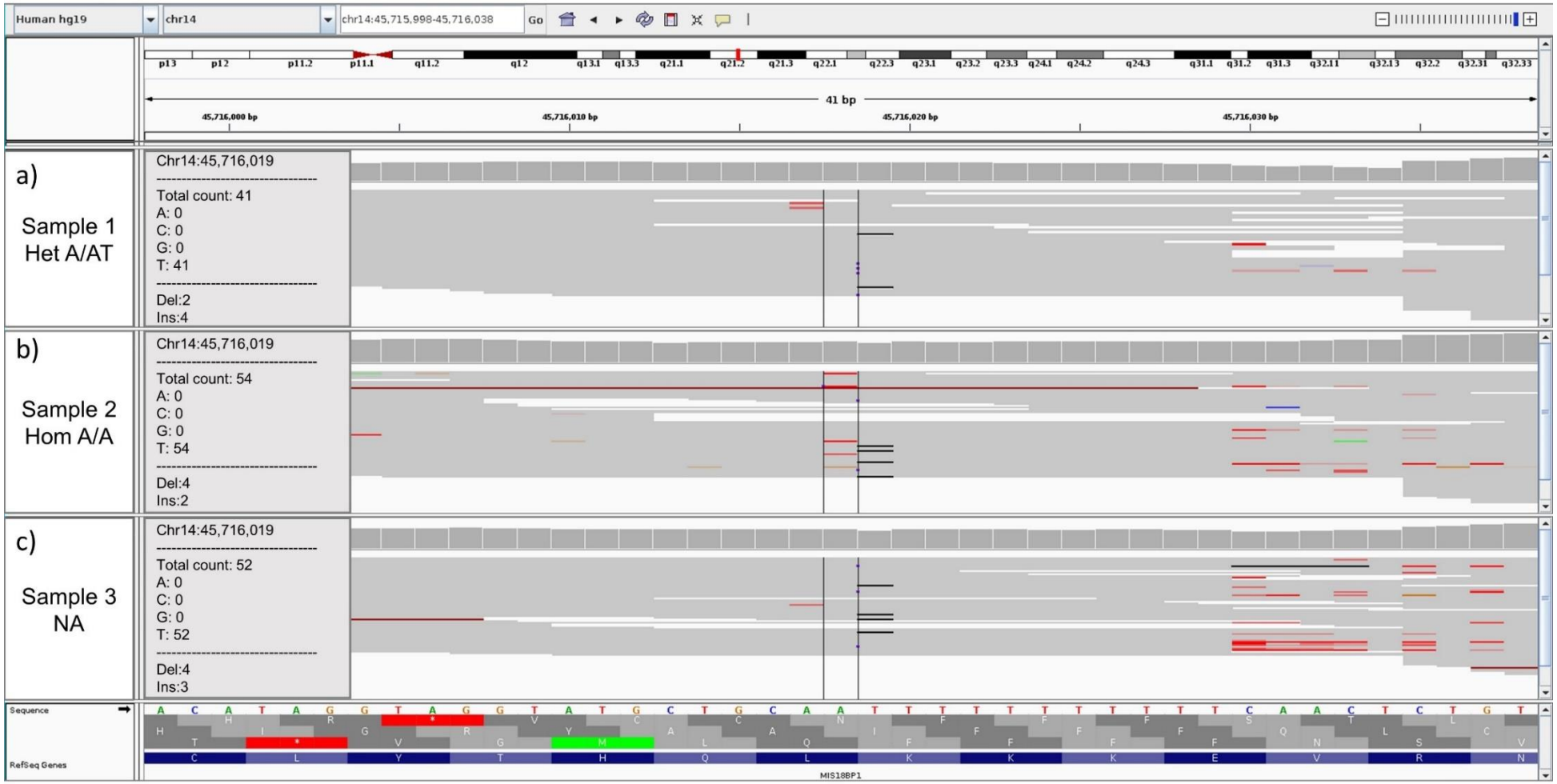


Figure 6.4: BAM file showing the frameshift insertion called in the Montreal set in gene *MIS18BP1* (rs546807245) a) Sample 1 was called heterozygous for the variant, b) Sample 2 was called as homozygous for the reference allele c) Sample 3 was unable to call the variant or it was filtered out due to poor quality.

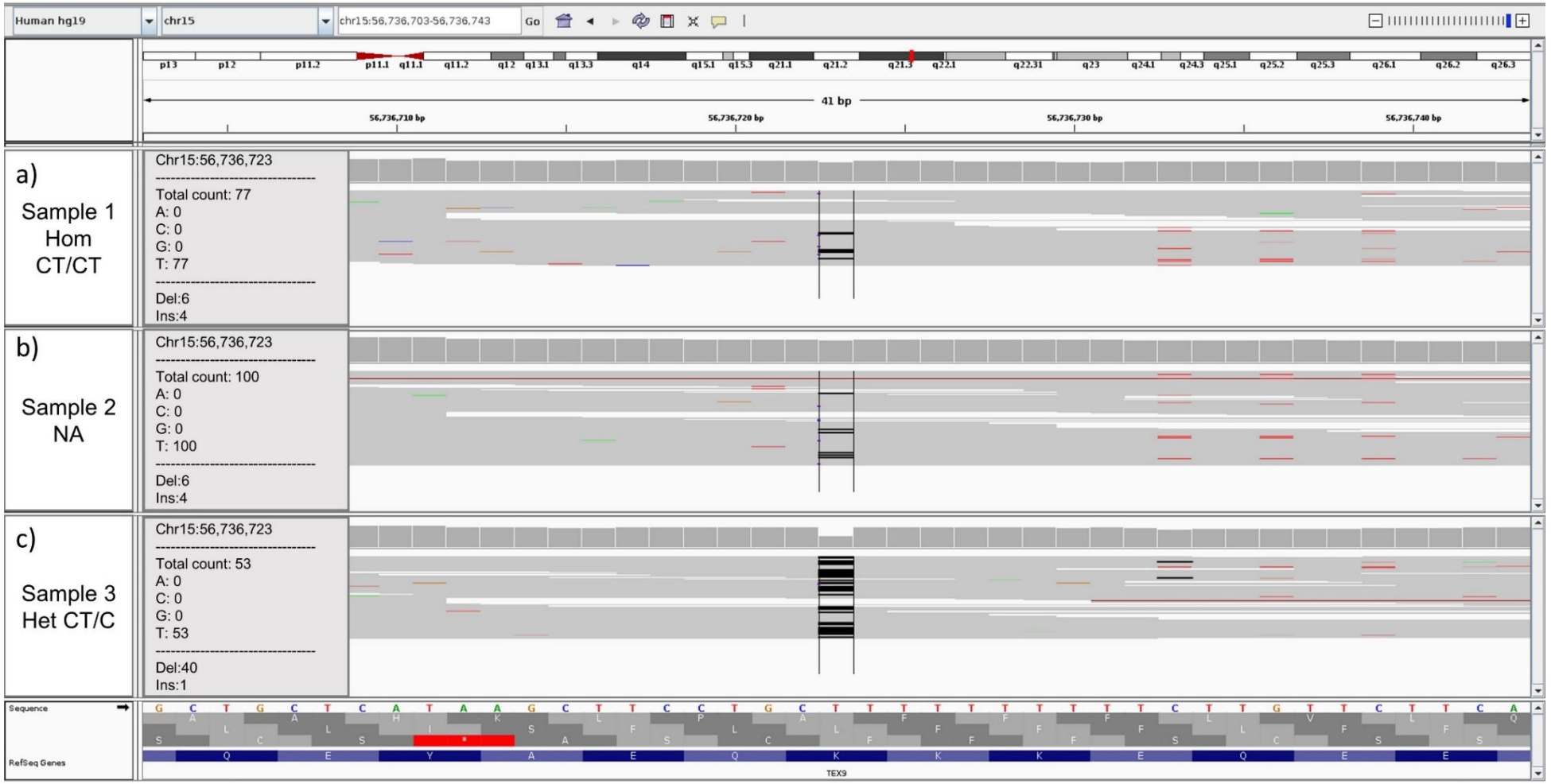


Figure 6.5: BAM file showing the frameshift deletion called in the Montreal set in gene *MNS1* (rs549395315) a) Sample 1 was called as homozygous for the reference allele, b) Sample 2 was unable to call the variant or it was filtered out due to poor quality, c) Sample 3 was called heterozygous for the variant.

CHROM	POS	REF	ALT	SYMBOL	Consequence	Existing variation	Montreal AC	GO ID	AF 1000 genomes	AF 1000 genomes NFE
19	1795984	G	A	ATP8B3	stop gained	rs187576612	1	GO:0000287	0.000998	0.003
11	104825726	C	CTTTT	CASP4	frameshift variant	rs555193787	1	GO:0004197	0.001398	0.005
2	202122956	T	C	CASP8	start lost	rs3769824	6	GO:0004197	0.034545	0.0467
13	100518634	C	T	CLYBL	stop gained	rs41281112	7	GO:0000287	0.021965	0.0268
17	78396004	A	G	ENDOV	splice acceptor variant	rs41298712	10	GO:0000287, GO:0003727	0.047125	0.0368
2	163124596	C	T	IFIH1	splice donor variant	rs35732034	1	GO:0003727	0.00619	0.0089
2	163134090	C	A	IFIH1	stop gained	NA	1	GO:0003727	0.001398	0.004
2	163136505	C	G	IFIH1	splice donor variant	rs35337543	5	GO:0003727	0.005591	0.0189
1	62676247	GAACT	G	L1TD1	frameshift variant	rs202029696	2	GO:0003727	0.005791	0.0179
16	77759403	T	A	NUDT7	stop gained	rs200408443	2	GO:0000287	0.000399	0.002
1	64089432	G	A	PGM1	splice donor variant	rs77043134	2	GO:0000287	0.003594	0.0139
17	74383107	CG	C	SPHK1	frameshift variant	rs549579958	2	GO:0000287	0.002995	0.0089
18	51880889	G	A	STARD6	stop gained	rs17292725	8	GO:0015485	0.014577	0.0328
6	41010518	G	A	TSPO2	splice donor variant	rs41273356	5	GO:0015485	0.008387	0.0278

Table 6.2: Variants in the Montreal set that were in enriched gene ontology terms in the Montreal set in comparison to 1000 genomes controls and in-house controls (FDR corrected p value < 0.01)

The gene *IFIH1* contains three different loss of function variants within the term 'single-stranded RNA binding' (GO:0003727), two splice donor variants (NM_022168.3: c.2807+1G>A, rs35732034 and NM_022168.3: c.1641+1G>C, rs35337543) and one stop gain (NM_022168.3: c.1879C>A, p.Glu627*). This term also contained a splice acceptor variant within the endonuclease gene *ENDOV* (NM_001164637.2: c.229-2A>G, rs41298712). This *ENDOV* variant was shared with the enriched 'magnesium ion binding' (GO:0000287) term. This term contained the largest number of different loss of function variants, with stop gains in *ATP8B3* (NM_001178002.2: c.1804C>T, p.Arg649*, rs187576612), *CLYBL* (NM_206808.3: c.775C>T, p.Arg259*, rs41281112), and *NUDT7* (NM_001105663.2: c.111T>A, p.Tyr37*, rs200408443), a splice donor variant in *PGM1* (NM_001172818.1: c.300+1G>A, rs77043134), and a frameshift deletion in *SPHK1* (NM_182965.2: c.854delG, p.Arg285Leufs, rs549579958). The term 'cysteine-type endopeptidase activity' (GO:0004197), was enriched due to the presence of two different loss of function variants in caspase apoptotic genes. A frameshift deletion was seen in *CASP4* (NM_001225.3: c.9_10insAAAA, p.Gly4Lysfs, rs555193787) and a start loss variant was identified in *CASP8* (NM_001080125.1: c.2T>C, p.Met1Thr, rs3769824). The cholesterol binding term was also only enriched due to the presence of two different variants, a stop gain in *STARD6* (NM_139171.1: c.55C>T, p.Arg19*, rs17292725) and a splice donor variant in *TSPO2* (NM_001010873.2: c.-21+1G>A, rs41273356).

The same GO enrichment analysis was performed on the small Ashkenazi Jewish population within the Montreal breast cancer set comparing to both the 1000 genomes and in-house control sets. Of the eight terms enriched in this set in comparison to the 1000 genomes controls, six were also significantly enriched in comparison to the in-house set. These six terms are 'cholesterol binding' (GO:0015485), 'integral component of lumenal side of endoplasmic reticulum membrane' (GO:0071556), 'DNA recombination' (GO:0006310), 'base-excision repair' (GO:0006284), 'peroxidase activity' (GO:0004601), and 'metalloaminopeptidase activity' (GO:0070006). Combined variant counts within the base-excision repair term were particularly low in control sets, with counts of 41 in 1000 genomes and 1 in the in-house set, in comparison to an allele count of 6 in the Ashkenazi Jewish set. Within these enriched terms, peroxidase activity related gene *EPX* and DNA recombination gene *RAD52* both contain two different rare loss of function variants.

6.5.3 WECARE gene prioritisation and breast cancer predisposition gene interaction analysis

Within the WECARE set, 216 genes were prioritised for targeted sequencing in a larger cohort using burden and variance based SKAT association testing and a proportional odds logistic regression. The proportional odds logistic regression test identified 94 genes that passed a significance threshold with a p value of less than 0.05 for a greater enrichment of variants in CBC individuals than UBC and NFE controls consecutively. The burden and variance based SKAT tests highlighted 92 genes each that were significantly enriched for variants within CBC individuals in comparison to UBC individuals. A set of

54 genes were shown to be associated with CBC by more than one of the three tests, and so were selected for further exploration via targeted sequencing.

The 216 prioritised genes were explored for an interaction with known breast cancer predisposition genes (*BRCA1, BRCA2, TP53, PTEN, CDH1, STK11, NF1, PALB2, ATM, CHEK2, NBN,* and *RECQL*) using the Cytoscape GeneMania plugin (Montojo *et al.*, 2010). Of these genes, 77 interacted either physically or via a related pathway with the input known breast cancer predisposition genes (figure 6.6). Two of the known breast cancer predisposition genes were within the prioritised gene list. The Neurofibromatosis Type 1 gene *NF1* was one of eight genes with a p value of less than 0.05 for an enrichment of variants within CBC individuals via all three statistical tests. It carried one loss of function and two different predicted deleterious or damaging missense variants within the WECARE samples. DNA damage checkpoint gene *CHEK2* was associated with breast cancer by burden based SKAT and proportional odds logistic regression.

Of the other prioritised genes that interacted with known breast cancer predisposition genes, those that directly interacted with BC genes were further explored. Excision repair gene *ERCC6* was shown to physically interact with both *CHEK2* and *TP53*. *ERCC6* was significantly enriched for variants in CBC individuals by burden based SKAT. One loss of function and four different deleterious and damaging missense variants were identified in this gene in the WECARE set. Further analysis of ExAC non-TCGA control set revealed that *ERCC6* is not commonly mutated in a healthy population, with only 47 occurrences of loss of function variants identified in non-Finnish Europeans from this cohort. In contrast, *BRD7* interacted with both *BRCA1* and *TP53*, however only carried one missense variant in the WECARE cohort and so was not considered a strong candidate.

Candidate gene *FOXM1* was shown to have a pathway interaction with *CHEK2*. This gene was highlighted as significantly enriched for variants in CBC individuals by proportional odds logistic regression and Burden based SKAT. *FOXM1* carried one loss of function variant and five different deleterious and damaging missense variants in the WECARE population. Only 18 occurrences of loss of function variants in this gene were identified in the NFE ExAC non-TCGA control set.

Another interesting candidate identified was the solute carrier gene *SLC9A3R1*. Identified as enriched for variants within CBC individuals by burden based SKAT, this gene shared a pathway interaction with *TP53*. One loss of function and four different deleterious and damaging missense variants were identified in this gene in the WECARE set, whereas only 34 loss of function variants were identified in the NFE ExAC non-TCGA control set. A literature search of this gene indicated that it may stimulate autophagy in breast cancer cells, and somatic variants have been seen in 25.8% of ovarian cancers within one study (Kreimann *et al.*, 2015; Liu *et al.*, 2015).

Another candidate gene, *STK11IP*, codes for an STK11 interacting protein and therefore was identified as having both a physical and pathway interaction with risk gene *STK11*. The gene had an enrichment

of variants in CBC individuals by burden based SKAT. This gene has a large number of loss of function variants in non-Finnish European ExAC non-TCGA set. However, of those 52,384 loss of function occurrences, only 83 are rare with an AF of less than 0.01, suggesting that this high number may be driven by a smaller number of common variants. This gene has one loss of function variant and three different deleterious and damaging missense variants in the WECARE set.

Variants within the candidate genes *NF1, CHEK2, ERCC6, FOXM1, SLC9A3R1,* and *STK11IP* can be seen in table 6.3. No variants were identified in the WECARE set in the Montreal GO based candidate genes. When the above WECARE candidate genes were queried in the Montreal set three variants were identified. A stop gain was found in *STK11IP* (NM_052902.3: c.718C>T, p.Arg240*, rs201204373) in two unrelated Ashkenazi Jewish individuals; this variant was not seen in the 1000 genomes control set. A missense variant was seen in one Ashkenazi Jewish individual in *ERCC6* (NM_000124.2: c.1996C>T, p.Arg666Cys, rs61760163), and a pathogenic (by ClinSig) frameshift deletion was identified in one sample in *CHEK2* (NM_001005735.1: c.247delC, p.Gln83Lysfs, rs587782766).

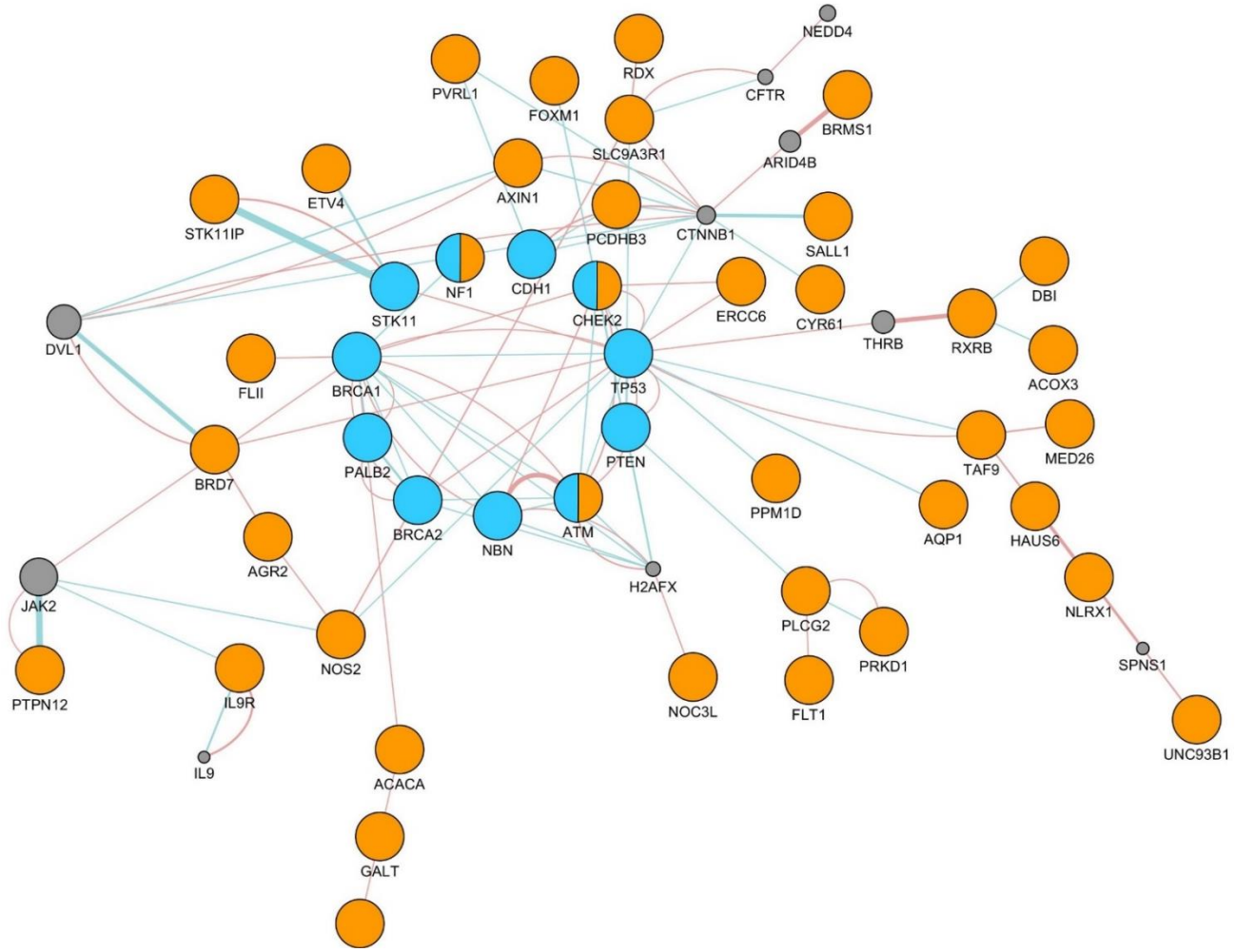


Figure 6.6: Cytoscape GeneMania interaction cluster of the 77 candidate genes (orange nodes) that interact physically (red line) or via a pathway interaction (blue line) with known cancer genes (blue nodes). Known cancer genes *NF1*, *ATM* and *CHEK2* were also candidate genes (half blue, half orange)

CHROM	POS	REF	ALT	SYMBOL	Consequence	Existing variation	CLINSIG	AC in CBC	AC in UBC	AC in NFE
2	220462640	G	T	STK11IP	start lost	rs681747	NA	30	12	19
2	220470710	G	A	STK11IP	missense variant	rs149218768	NA	3	1	0
2	220473373	T	C	STK11IP	missense variant	rs765799088	NA	1	0	0
2	220473374	T	G	STK11IP	missense variant	rs753147630	NA	1	0	0
10	50678356	A	C	ERCC6	missense variant	rs61760166	NA	2	0	2
10	50681659	T	G	ERCC6	intron variant	rs4253196	pathogenic	1	0	0
10	50682279	C	T	ERCC6	missense variant	rs751703364	NA	1	0	0
10	50690906	G	A	ERCC6	missense variant	rs61760163	pathogenic	1	0	2
10	50701164	T	A	ERCC6	missense variant	rs200832611	NA	1	0	0
10	50713929	C	A	ERCC6	splice donor variant	rs371739894	pathogenic	1	1	0
12	2968078	G	A	FOXM1	missense variant	rs28919870	NA	9	3	5
12	2968523	G	A	FOXM1	missense variant	rs151053319	NA	1	0	0
12	2970472	CAAAG	C	FOXM1	frameshift variant	rs773729390	NA	2	0	0
12	2973547	G	T	FOXM1	missense variant	rs28990715	NA	12	9	9
12	2973861	G	T	FOXM1	missense variant	rs552495230	NA	1	0	0
17	29497015	G	A	NF1	missense variant	rs876659079	uncertain significance	1	0	0
17	29684326	C	T	NF1	stop gained	rs786201367	pathogenic	1	0	0
17	29687560	T	C	NF1	missense variant	rs752541243	uncertain significance	1	0	0
17	72758167	G	A	SLC9A3R1	missense variant	rs41282065	pathogenic	3	1	0
17	72758211	C	A	SLC9A3R1	missense variant	rs373243340	NA	1	0	0
17	72758239	A	G	SLC9A3R1	missense variant	NA	NA	1	0	0
17	72758301	C	T	SLC9A3R1	missense variant	NA	NA	1	0	0
17	72764377	T	C	SLC9A3R1	splice donor variant	rs138547261	NA	1	0	0
22	29091226	TA	T	CHEK2	frameshift variant	rs587780174	pathogenic	1	0	0
22	29091856	AG	A	CHEK2	frameshift variant	rs555607708	pathogenic	4	3	0
22	29092947	C	T	CHEK2	missense variant	rs730881688	uncertain significance	1	0	0
22	29095854	T	C	CHEK2	missense variant	rs587780194	uncertain significance	1	0	0
22	29107962	A	G	CHEK2	missense variant	rs141776984	uncertain significance	0	1	0
22	29121018	C	T	CHEK2	missense variant	rs137853009	pathogenic	1	0	1
22	29121087	A	G	CHEK2	missense variant	rs17879961	pathogenic	3	0	0
22	29121242	G	A	CHEK2	missense variant	rs137853007	pathogenic	2	0	0
22	29121326	T	C	CHEK2	missense variant	rs28909982	likely pathogenic	0	1	0
22	29130520	C	T	CHEK2	missense variant	rs141568342	likely pathogenic	1	0	0
22	29130540	G	A	CHEK2	missense variant	rs730881695	uncertain significance	1	0	0

Table 6.3: Variants identified in candidate genes in the WECARE set, including allele counts in contralateral breast cancer (CBC), unilateral breast cancer (UBC) and non-Finnish European Female controls (NFE)

6.6 Discussion

The Montreal and WECARE breast cancer sets both utilise the collection of participants who are at greater risk of hereditary breast cancer. Within the Montreal set, this is via the selection of individuals with a strong family history of breast cancer and those from the homogenous Ashkenazi Jewish population. The potential to identify founder variants within this set gives an advantage in identifying rare variants that might not be seen in significant numbers in a more heterogenous population. The selection of individuals with multiple primary tumours within the WECARE set, and comparison to those with single breast cancers also enriches the case population for a potential hereditary risk of breast cancer (Cybulski, Nazarali and Narod, 2014). Importantly, care was taken to ensure that the CBC individuals were asynchronous, and so the second tumour was likely to be a primary and not a result of metastasis.

The Montreal set was around a tenth the size of the WECARE set and therefore required a more tailored approach to identify candidate variants. The use of GO enrichment analysis was developed to aggregate variant counts across the set and test for an enrichment of variants in each tested term. In smaller sets where the identification of rare variants in more than one individual is unlikely, aggregation of variants into genes increases the likelihood of an association being identified. Aggregation of variants into ontology terms further increases this likelihood. This technique draws functional links between variants that may indicate a shared molecular pathway or biological process and therefore potentially may produce a shared phenotype.

A number of GO terms were identified as enriched with variants in the Montreal set in comparison to both 1000 genomes control set and an in-house generated set. Using the 1000 genomes as a control set allows for the selection of terms that are significantly enriched in breast cancer cases in comparison to healthy individuals. However, using external datasets for association testing is likely generate false positive results due to technical and ethnic differences. Within the WECARE set analysis, an ethnically comparable subset of 1000 genomes data were downloaded and analysed to reduce technical and ethnic differences. Within the GO analysis, to minimise the effect of ethnic differences tests were only run on the European subset of 1000 genomes. Although technical differences can be reduced slightly by performing comparable genotype filters to remove low quality genotypes, filtering differences upstream of this, in addition to other technical differences, may still cause confounding results. By comparing to an in-house set, enriched ontology terms in cases in comparison to other (non-breast) cancer controls can be selected where both sets have been analysed using almost identical sequencing conditions. However, these controls are not healthy individuals, and as many breast cancer genes are also linked to other cancers, these cannot be treated as pure controls and so cannot be used as the sole comparator. However, as they were sequenced at the same facility, and ran through the same VCF generation pipeline and filtering, there will be fewer technical differences. Therefore, where terms are enriched for

variants in comparison to both control sets, the statistical significance is unlikely to be due to technical differences alone.

Four molecular functions were found to be enriched for loss of function variants in the Montreal set in comparison to 1000 genomes individuals and in-house controls. The term cholesterol binding was also associated with the Ashkenazi Jewish set when tested independently. However, this set was only comprised of two different variants, both with high allele counts. Although these variants passed rarity filters set using 1000 genomes Europeans, they were each present in ~3% of ExAC non-TCGA NFE controls and so are relatively common in a second healthy population.

In contrast, the terms 'single-stranded RNA binding' (GO:0003727) and 'magnesium ion binding' (GO:0000287) contained five and six different variants respectively. Single stranded RNA binding has previously been associated with breast cancer survival, with mRNA transporter *RBM47* being identified as a suppressor of breast cancer progression (Upadhyay *et al.*, 2013; Vanharanta *et al.*, 2014). However within this set, variants are predominantly in the interferon induced gene *IFIH1*, variants within which are predominantly associated with encephalopathy in Aicardi-Goutières syndrome (Oda *et al.*, 2014).

Of the six genes with variants in the 'magnesium ion binding' (GO:0000287) term, the majority have no previous association with breast cancer. However sphingosine phosphorylation gene *SPHK1* has been shown to be upregulated in breast cancer, with high expression correlating with poor survival and treatment response (Datta *et al.*, 2014; Zhang *et al.*, 2014; Maczis, Milstien and Spiegel, 2016). The variant identified in this cohort is a loss of function frameshift deletion of one base pair and so is unlikely to cause an overexpression or be linked to the described phenotype in this case.

The majority of breast cancer predisposition genes currently identified function in DNA repair checkpoints and homologous recombination. It is therefore reasonable to hypothesise that variants in DNA repair related GO terms might also be strong candidates for a novel association with breast cancer risk. Genes that interact with the DNA repair pathway but are not currently known to be DNA repair genes would also be strong candidates but would be missed by this GO enrichment analysis. As the participants included in this study previously tested negative for known breast cancer genes, it is unsurprising that DNA repair terms were not identified as enriched via this analysis. The identified candidate *SLC9A3R1*, is not involved in DNA repair via known mechanisms. However, tools such as GeneMania can be used to identify genes that could interact with this pathway, such as *SLC9A3R1* which physically interacts with *TP53*. This solute carrier gene has been shown to stimulate autophagy via the stabilization of BECN1 in breast cancer cells (Liu *et al.*, 2015). Defects in the autophagy process have been shown to associated with increased tumorigenesis, and a somatic loss of *BECN1* has been seen in breast, ovarian, and prostate cancers (Aita *et al.*, 1999; Liang *et al.*, 1999).

The transcription factor Forkhead Box M1 (encoded by *FOXM1*), was first associated with human cancer by Teh et al, who discovered upregulation of the gene in basal cell carcinoma (Teh *et al.*, 2002).

It has since been shown that high expression of this gene is an early progressor of tumour cells, providing genomic instability and allowing the accumulation of further DNA damage (Gemenetzidis *et al.*, 2010; Jia *et al.*, 2010; Teh *et al.*, 2010). The missense variants identified in the WECARE set in *FOXM1* could affect gene regulation, however the frameshift deletion in this gene is unlikely to be linked to an upregulation in cancer cells.

Germline variants in serine-threonine protein kinase 11 gene (*STK11*) were identified as the cause of Peutz-Jeghers syndrome in 1998 (Hemminki *et al.*, 1998). Affected individuals often have multiple hamartomatous polyps but also carry a greater risk of breast, colorectal, small bowel, pancreatic, gastric, and ovarian cancer (Volikos *et al.*, 2006). Currently there is limited research into the STK11 interacting protein STK11IP, however within this study it has been suggested as a new breast cancer candidate gene in two independently collected datasets. Within the WECARE set multiple missense variants as well as a start-loss variant were seen in a number of individuals. Additionally, a rare stop- gain variant (NM_052902.3: c.718C>T, p.Arg240*,, rs201204373) was identified in two unrelated Ashkenazi Jewish individuals within the Montreal set. This rare variant could represent a founder variant in the homogenous population and so may have not been previously identifiable in heterogenous population studies.

6.7 Summary

This study explores predisposition to breast cancer through the analysis of different datasets that utilise homogenous and high-risk patient populations. Variant aggregation into genes and ontology terms was used in both sets to provide greater statistical power to association tests. Variants involved in single-stranded RNA binding and magnesium ion binding have been recommended as potential candidates for further exploration in larger sets. A number of candidate genes from the WECARE study have been put forward for further study. In particular, *STK11IP*, was identified in both breast cancer sets. The presence of multiple occurrences of a truncating variant within this gene in an Ashkenazi Jewish population could be indicative of a founder variant. These studies provide techniques for analysing underpowered sequencing data. In particular they emphasise the need to cross reference results between related datasets. This provides the opportunity to test candidates under different sample selection criteria and often different sequencing and analysis environments. Identifying candidates within two different datasets provides a more robust argument for that candidate gene's involvement in disease predisposition.

7 Discussion

The primary focus of this work has been to identify genes that when altered predispose to various cancer types, looking specifically at early onset and high-risk cases without known genetic risk factors. In studies of common cancers and diseases, larger datasets are easier to obtain, providing increased statistical power when looking for risk factors. However gathering large sample sets of rare disease affected individuals can prove far more difficult and often statistical tests on the resulting smaller sets are underpowered to detect rare variants (Hoffmann, Marini and Witte, 2010). By sequencing multiple individuals from families with a strong cancer history and individuals from homogenous populations, the frequency of rare variants is increased in comparison to an unrelated population. Therefore, in such sets the likelihood of identifying rare candidate variants for rare cancer syndromes is increased. Where such groups are not available, selecting individuals with an early age of onset or those with multiple primary cancers increases the likelihood that the disease is caused by hereditary genetic factors.

Careful study design and sample selection has enabled the production of datasets within this study that are enriched for genetic risk factors. Sequencing data from multiple affected and unaffected family members and subsequent segregation analysis can show whether the identified genotype segregates with the phenotype. This has been particularly important where a literature review of candidate genes reveals no clear association between the affected pathway and the phenotype such as *MYH9* variants in MALTA syndrome. As pathogenic variants in *MYH9* are already linked to a well described blood disorder (Althaus and Greinacher, 2009), it originally appeared unlikely that protein-affecting variants in this gene would also prove causal to a distinct skin tumour. However further segregation analysis in a number of additional families provided clear evidence that *MYH9* was a strong candidate gene.

The study of sequencing data from a homogenous population of Ashkenazi Jewish heritage allowed for the identification of a rare, potential founder variant in more than one unrelated individual that may increase breast cancer risk. The gene *STK11IP,* an interacting partner of Peutz-Jeghers syndrome gene *STK11,* was identified as a potential candidate for CBC predisposition in the WECARE breast cancer cohort, prior to the identification of additional variants in the Montreal breast cancer set. The differences in the two cohorts, both in terms of study design and ethnicity, meant that they required separate analysis. Variant aggregation into genes and statistical prioritisation could be applied to the larger WECARE set to select candidate genes for further targeted sequencing. Whereas the smaller Montreal set benefited from a GO enrichment analysis, further aggregating variants into terms to increase statistical power, which led to the identification of two different rare, loss of function variants in DNA recombination and repair related gene *RAD52* in the Ashkenazi Jewish cohort.

This study has also benefited from detailed clinical and tumour histology data. Within the ACC set, the histological subtype of the tumours was reviewed by an expert pathologist (Dr Alison Marker) and tumour information was provided including size and weight, stage at diagnosis and mismatch repair

patterns. Such detailed information is not always readily available when using external datasets. However cohort publications on TCGA subsets often provide in depth clinical, tumour, and ethnicity information, allowing for subtype specific analysis comparable to work completed on in-house data (Bass *et al.*, 2014; Zheng *et al.*, 2016). This makes TCGA data a useful resource for integration into local analysis to explore candidate variants or increase sample numbers.

Even with an increase in sample numbers provided by the integration of external data, a rare disease dataset often remains statistically underpowered to detect rare genetic variants (Hoffmann, Marini and Witte, 2010). Calculations can be made to estimate the power of a given statistical test, which can then be used to predict the sample size required to reach the desired statistical power (Sham and Purcell, 2014; Wang *et al.*, 2014). However, as previously stated, the recruitment of these rare disease affected individuals is the limiting factor in the study design process. To partially overcome the difficulties of low statistical power, the analyses described here have made extensive use of variant aggregation into genes and functionally related gene groups. Tools such as the GeneMania Cytoscape plugin provide information about protein interactions by implementing the GeneMania algorithm (Mostafavi *et al.*, 2008). The tool uses interaction data from a number of sources including BIOGRID (Chatr-Aryamontri *et al.*, 2015), GEO (Barrett *et al.*, 2009), I2D (Brown and Jurisica, 2005) and Pathway Commons (http://www.pathwaycommons.org). Protein networks are drawn via co-expression, co-localization, genetic interactions, physical interactions, pathway interactions, predicted interactions, and shared protein domains. The usage of a variety of heterogenous data by GeneMania provides interactions that might be missed in a literature search or using pathway data such as the Kyoto encyclopedia of genes and genomes (KEGG) (https://www.kegg.jp/).

One of the limitations of GeneMania is that it provides no information as to the function of the clustered genes. For example, *PALB2* and *BRCA2* may be described by the tool as physically interacting, however users must draw their own conclusions about the involvement of both genes in DNA repair. Within this study, GO terms were applied to identified GeneMania interaction clusters, providing a description of the biological process within which those genes may be interacting. Within this thesis, the method of testing for an enrichment of variants within a biological process was first used in the HDGC set, supporting results in the literature describing a number of HDGC cases with DNA repair variants (Hansford *et al.*, 2015; Sahasrabudhe *et al.*, 2017; Slavin *et al.*, 2017). The use of GO terms in addition to GeneMania allowed for the study and identification of variants in potential candidate genes associated with the DNA repair ontology term or interacting with known DNA repair genes.

GO enrichment analysis is a well-used form of variant aggregation and has been described in multiple studies including identifying functionally related differentially expressed gene clusters, identifying gene sets that are haploinsufficient, and identifying processes that are often targeted by somatic variants in neuroblastomas (Wang *et al.*, 2006; Dang *et al.*, 2008; Sanders *et al.*, 2013). Other sets within this study

used GO terms independently of GeneMania clustering. This analysis was utilised in the breast cancer cohort due to the larger sample sizes and therefore larger numbers of variant carrying genes in comparison to other datasets within this study. Larger networks of genes intuitively had greater numbers of interactions within GeneMania and so did not generate functionally informative clusters. Where clusters could not be drawn, all filtered GO terms were tested for an enrichment of loss of function variants. Although this did not have the benefit of analysing genes highlighted as interacting by multiple sources, it still provided informative results about biological processes enriched for rare, loss of function variants in case sets.

Analysis of both the ACC and HDGC cohorts resulted in the identification of loss of function and likely disease predisposing variants in DNA repair related candidate genes. Within HDGC, breast cancer gene *PALB2* and Lynch syndrome gene *MSH2* were identified with rare loss of function variants. Tumour immunohistochemistry also revealed a somatic loss of Lynch syndrome proteins MSH2 and MSH6 within two ACC tumours from different individuals, with one individual also carrying a germline loss of function *MSH6* variant. No clear difference was seen in the predicted loss of function consequences of *PALB2, MSH2,* and *MSH6* variants identified in these cohorts in comparison to those pathogenic variants seen in the more common *PALB2* and Lynch syndrome phenotypes of breast and colorectal cancer respectively, nor was there any obvious genotype-phenotype correlation within the limited number of pathogenic variants identified. The HDGC family carrying a loss of function *PALB2* variant also had a history of breast cancer on both maternal and paternal sides, in addition to gastric cancer, lung cancer, and laryngeal cancer on the paternal side. Despite this range of cancer phenotypes, three of the six siblings in the tested generation were diagnosed with diffuse gastric cancer before the age of 60, suggesting an increased risk for this specific phenotype within this generation. With no sequencing data available from the previous familial generation, one can only speculate that the other identified cancers could be influenced by environmental factors, combinations of different genetic factors, or be simply due to chance. It is possible that pathogenic variants in DNA repair genes such as *PALB2* may act as general low-level cancer predisposing genes beyond their role as higher penetrance genes in certain cancer types. This concept was reinforced by a recent study of TCGA data showing pathogenic *BRCA1* and *BRCA2* variants in 25 different cancer types (Huang *et al.,* 2018).

This Huang et al pan-cancer study also identified rare CNVs in cancer genes *BRCA1, FH, MSH6, NF1, PALB2,* and *PTEN* (Huang *et al.,* 2018). The exome hidden Markov model (XHMM) used to identify these CNVs was similarly used in analyses described within this thesis. CNV analysis was undertaken in families with a very strong history of early onset disease but no identifiable pathogenic germline variants, however no clinically relevant alterations were discovered. Within exome data, CNV detection relies on depth of coverage over the covered exons in comparison to breakpoint focused analysis such as BreakDancer (Chen *et al.,* 2009), which can be used to detect structural variant breakpoints across genomic data. In addition to using the XHMM algorithm on WES data, Huang et al identified half of

the reported variants using genotyping SNP-array data. Of the more than 95,000 CNVs detected, less than 4,000 were consistent across the two study types, largely due to the genotyping data covering intronic regions (Huang *et al.*, 2018). A combinatory approach to identifying CNVs using different data sources from the same sample improves the likelihood that variants called are not a result of variation in amplification efficiency and sequencing depth. The XHMM algorithm normalises read depth data across samples to account for poorly covered regions across the sequencing library (Fromer and Purcell, 2014). However, within this study, a number of identified CNVs called using this algorithm could not be validated. To improve identification of CNVs and other structural variants, a number of studies, including research into structural variants in 1000 genomes phase 3 data, used low coverage WGS data to increase the range of detectable structural variants (Sudmant *et al.*, 2015). This can be used in conjunction with more affordable WES data at a higher coverage to detect structural variants including deletions, insertions, duplications, and inversions, alongside single nucleotide variants and short insertions and deletions (Tattini, D'Aurizio and Magi, 2015).

A number of other candidate genes have been identified by work presented in this thesis which have not previously been described as affecting predisposition to cancer or tumour development. One key example of this was the identification of protein-affecting variants in myosin gene *MYH9* in MALTA syndrome. The literature currently describes no instances of this gene being involved in cell proliferation or the development of sweat ducts. Unlike many known cancer predisposition genes which affect cell proliferation in a number of different tissue types, the two phenotypes associated with *MYH9* pathogenic variants are distinct in terms of their cellular effects, with one influencing cell size and the other influencing proliferation. Functional work (described in more detail below) would be required to fully understand the many roles of this non-muscle myosin in different tissue types and how pathogenic variants in different regions can cause two distinct phenotypes.

Where available, tumour studies were undertaken to provide additional insight into the role candidate variants could be playing in disease development. Within the set of MALTA families, only one tumour block was available where sufficient DNA could be extracted to perform additional somatic sequencing. This limited the ability to determine whether identified somatic variants were unique in this individual or common and relevant to the tumour type. Within the HDGC and ACC cohorts, MSI and IHC were used to determine if tumours were deficient in mismatch repair. Although such a deficiency could suggest the presence of a germline Lynch syndrome variant, this is not always the case, as was shown in one individual whose ACC tumour showed a loss of MSH2 and MSH6, but who carried no identifiable exonic germline mismatch repair variant (even after a manual search for variants of all consequence in mismatch repair genes). In the case of HDGC, IHC and MSI were critical in showing that the identified germline *MSH2* loss of function variants were not causing a complete deficiency in mismatch repair. This tumour information can therefore be used both to ascertain whether a germline

variant is having the predicted effect on protein function, and to point to a possible means for disease predisposition.

Both somatic variants and expression data can be used to identify pathways implicated in tumorigenesis. Where somatic data show passenger variants and driver variants which have been selected for by the tumour microenvironment, RNA sequencing data show the resulting gene expression patterns which can elucidate the role of both somatic and germline protein-affecting variants. Within the TCGA-ACC cohort, a change of expression of both *PDE2A* and *PKIB* suggest that the cAMP-PKA pathway is implicated in the development of the oncocytic ACC subtype. These expression changes could be caused by many different somatic or methylation events changing regulation of the cAMP-PKA pathway; however, it is the end result, the overactivation of this pathway, that is relevant to future research. Expression data can therefore in some senses be more informative than somatic data for detecting affected pathways and suggesting potential therapeutic targets.

Although this study has provided insight into rare cancer and disease predisposition, the question of missing heritability remains. Particularly in the case of the ACC subset, young onset cases (including one described individual diagnosed as young as 18 years of age) are being identified with no obvious disease cause, either genetic or environmental. A study by Tomasetti and Vogelstein proposed that chance effect is underestimated in disease development, and that the majority of cancers are driven by random somatic variants occurring during DNA replication (Tomasetti and Vogelstein, 2015b). However this study was rather controversial, with some countering its assertion that such "bad luck" driven malignancies would be resistant to preventative methods and others taking issue with the statistical methods used (Altenberg, 2015; Ashford *et al.*, 2015; Gotay, Dummer and Spinelli, 2015; O'Callaghan, 2015; Potter and Prentice, 2015; Song and Giovannucci, 2015; Tomasetti and Vogelstein, 2015a; Wild *et al.*, 2015). Others also noted that Tomasetti and Vogelstein did not allow for an interaction between extrinsic (such as UV damage) and intrinsic processes (such as DNA replication) during their analysis and that these extrinsic processes may cause the same process of damage accumulation via stem cell division (Wu *et al.*, 2016).

The "bad luck" theory may explain some of the missing heritability within this study, however there is also a possibility that risk variants were not identified due to limitations in current technology with incorrect predictions of variant pathogenicity resulting in the filtering out of pathogenic variants. Additionally, some cancer phenotypes could have been a result of the interaction of many different genetic factors. Some GWA studies have aimed to generate polygenic risk scores, with one study in breast cancer susceptibility describing SNPs that account for around 44% of familial relative risk (Michailidou *et al.*, 2017). Although polygenic risk scores have provided informative classifiers for complex diseases such as prostate cancer (Szulkin *et al.*, 2015), the statistical power required to define such scores is likely to be unachievable in many rare diseases. For such diseases, a similar polygenic

risk may be ascertainable by looking at smaller numbers of rare variants in highly conserved genes and exons. In particular it would be interesting to look at patterns of rare variant occurrence in pathway such as DNA damage repair across various cancer types. This could be further studied using large control sets to identify parts of the DNA repair pathways that are highly conserved, and to identify variant barcodes that are rare in healthy populations.

In addition to creating a greater understanding of disease predisposition, this study aimed to create and describe new analysis methods for next generation sequencing data and tools for use by researchers interested in studying a range of genetic data. The use of control data is imperative when interpreting the results of germline sequencing data to determine how often variants appear in a healthy population. The majority of control datasets have an online portal to allow researchers to query variation in genes and variants of interest. The 1000 genomes data do not have a specific browser but some data including variant population frequencies are accessible online through Ensembl (http://www.ensembl.org) (Auton *et al.*, 2015). The genome Aggregation Database (gnomAD) and the Exome Aggregation Consortium (ExAC), which combine sequencing data from a number of sources including TCGA, 1000 genomes, and Myocardial Infarction Genetics Consortium, have web portals which provide summary data for genes (including coverage metrics, CNV counts, and lists of identified variants) and variants of interest (showing population frequencies and IGV snapshots of the reads containing the variant in BAM files) (Lek *et al.*, 2016).

Although the ExAC/gnomAD browser format is extremely user friendly and informative, it does not provide users an opportunity to select or filter by data source and therefore remove data that is linked to the researcher's phenotype of interest; for example, within this study, data from TCGA were excluded. ExAC however does provide users an opportunity to download a VCF excluding variants from key psychiatric studies and from TCGA. As part of the analysis done in this thesis, a browser was created for ExAC non-TCGA data and 1000 genomes release 1 data to allow users to interrogate these VCFs in a user-friendly web app. This improves the reproducibility of this study as all results taken from the ExAC non-TCGA and 1000 genomes resources are publicly available from a web address (https://medgenbrowser.wordpress.com/). It also provides a useful resource for cancer-focussed genetic researchers who may not have the required bioinformatics skills to download, filter, and query a raw VCF file. At the time of writing, since the web address was published in March 2018, collectively the browsers have been used over 100 times by over 40 different users, suggesting that this is a resource that is of use to the wider genetics community.

The results of this study, including tools and data created, have been or will be shared with the genetic community through open-access publications, collaborations, and GitHub repositories. Novel next generation sequencing datasets are a valuable resource for further studies to build upon. Many studies now combine sequencing data from multiple sources to provide larger datasets for greater powered

statistical analysis (Rinella *et al.*, 2013; Hansford *et al.*, 2015). The work presented here has generated one of the largest completed sequencing studies of the oncocytic ACC subtype. Making data available for future studies drastically changes the scope of rare disease research, from what were once individual case studies to larger combined cohorts of published data. Additionally, the creation and sharing of tools and analysis techniques allows genetics research to be more accessible for the wider community and improves the reproducibility and validity of generated research.

8 Future directions

Work presented in this thesis could be expanded upon to provide a greater understanding of how identified candidate variants are influencing disease predisposition. The studies described here used small sample sizes and limited tumour sequencing data. For further genetic and functional analysis of variants identified in this study, both sample sizes and tumour analysis could be expanded where possible. In some cases, further work would involve an accumulation of available external data or the recruitment of additional samples. In sets where strong candidate genes have been suggested, functional work would help to understand how these variants are producing or affecting the identified phenotype.

8.1 Hereditary diffuse gastric cancer

Due to the established availability of preventative surgery for HDGC, the field would greatly benefit from a further understanding of disease predisposition. The identification of loss of function variants in *PALB2* within this study and external studies suggests that loss of function variants in this gene play a greater role in HDGC predisposition than previously thought. Data from this study has since been pooled with data from collaborators to create a larger HDGC WES study. This combined study may have the statistical power to quantify the lifetime risk of HDGC development in individuals with loss of function *PALB2* variants.

An interesting family identified within this study carried germline, loss of function variants in both *ATR* and *NBN*. Two of the siblings in this family had prophylactic gastrectomies and no tumour was available from the affected third sibling or either of the affected parents. Future plans for this work include the microdissection of single signet ring cells from tissue collected from the risk-reducing gastrectomies performed in two members of this family. If enough cells can be gathered for DNA extraction, Sanger sequencing can be performed to look for loss of heterozygosity in the two identified predisposition variants. Additionally, depending on the amount of DNA extracted from these cells, somatic WES could be used to look for signatures of DNA damage repair deficiency. It would be particularly interesting to see if loss of heterozygosity or patterns of somatic variants could be attributed to one of the variants in *ATR* and *NBN*, or to both of them.

8.2 MALTA syndrome

Despite extensive efforts, this study has been unable to clearly elucidate the dual role of non-muscle myosin gene *MYH9* in MALTA and MYH9-related platelet disorder (MRPD). To fully understand how

the two germline variant sets produce these two phenotypes, one would need to study how these variants differentially affect non-muscle myosin function. Work could be completed initially using cell lines, introducing variants implicated in either MALTA or MRPD to observe the differences in cell growth, migration, and proliferation rates. One could also mirror previous work by Hu et al (Hu, Wang and Sellers, 2002) who studied the kinetic effects of MRPD variants on myosin motor activity. By performing a comparable study using MALTA variants one could compare the ATPase activity and motility rate of non-muscle myosin IIA when affected with variants causing each phenotype.

In silico options for further study on this set include further studying the previously reported post-transcriptional modifications on *MYH9*. This idea was originally considered to explain how germline variants that appear to disrupt a key motor function do not affect many more tissues throughout the body, leading to the suggestion that interrupting a specific post-transcriptional modification pattern may affect one tissue type or one specific cellular role. Although the post-transcriptional modification sites surrounding these variants have been briefly explored, a full study into this would require the analysis of all reported variants in both MALTA and MRPD.

8.3 Adrenocortical Carcinoma

The rarity of ACC, specifically the oncocytic subtype, limited the analysis performed on this set to only a handful of samples. In spite of this, a number of predisposition genes were recommended for further study. One of the future steps on this project will involve the WES of FFPE material from a number of samples including two additional oncocytic samples, generating the largest WES cohort of oncocytic ACC generated so far. This study aims to further elucidate the role of mismatch repair and the cAMP-PKA pathway in tumour development of oncocytic ACC and to identify any distinguishing features of this rare subtype that may help in diagnosis or treatment. Additionally, this somatic sequencing will be analysed to look for loss of heterozygosity (LoH) of candidate variants suggested in the germline data presented here. If any candidates present with LoH, the gene containing these variants could be explored within other studies such as the tumour and normal sequencing data available from oncocytic ACC samples within the TCGA cohort. This could provide further evidence for a role for these candidates in oncocytic ACC predisposition.

8.4 Breast cancer predisposition

For the WECARE study of contralateral breast cancer predisposition, candidate genes were proposed for further targeted sequencing in a larger cohort. This included *STK11IP* which was identified with a potential founder variant in the Montreal Ashkenazi Jewish set. An Ampliseq panel of primers covering the coding sequences of candidate genes was designed and sequenced in a further set of germline DNA samples from around 500 additional women from the WECARE study with either CBC or UBC. This should provide greater statistical evidence to explain the role of candidate genes in disease predisposition.

9 Summary

This work has improved the knowledge of rare disease predisposition to enable more accurate disease management strategies. As well as identifying a greater role than previously thought for *PALB2* in HDGC predisposition, the combinatory effect of multiple cancer predisposition variants was explored, further adding to the suggestion that cancers should no longer be considered monogenic diseases. The new syndrome MALTA was delineated through the identification of rare variants in myosin gene *MYH9* in all identified cases. This warrants further exploration of the role of myosins in cellular proliferation with a particular focus on genotype-phenotype correlations which are likely to be important for this gene. In addition to the suggestion of possible new candidate genes for ACC predisposition, the cAMP-PKA pathway was identified as potentially playing a role in the development of the extremely rare oncocytic subtype. Candidate breast cancer predisposition genes were identified for further study, in particular *STK11IP* and its potential new founder variant. Finally, this study provided an opportunity to explore many genetic analysis techniques for the study of rare diseases. During this process novel tools such as the control browser web app were developed for use by the wider research community.

10 References

Adams, J. M. and Cory, S. (2007) 'The Bcl-2 apoptotic switch in cancer development and therapy', *Oncogene*, 26(9), pp. 1324–1337. doi: 10.1038/sj.onc.1210220.

Adzhubei, I., Jordan, D. M. and Sunyaev, S. R. (2015) *Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2, Current Protocols in Human Genetics*. doi: 10.1002/0471142905.hg0720s76.Predicting.

Aita, V. M. *et al.* (1999) 'Cloning and genomic organization of beclin 1, a candidate tumor suppressor gene on chromosome 17q21.', *Genomics*, 59(1), pp. 59–65. doi: 10.1006/geno.1999.5851.

Alexander, A. and Paulose, K. P. (1998) 'Oncocytic variant of adrenal carcinoma presenting as Cushing's syndrome.', *The Journal of the Association of Physicians of India*, 46(2), pp. 235–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11273124.

Allolio, B. and Fassnacht, M. (2006) 'Adrenocortical Carcinoma: Clinical Update', *The Journal of Clinical Endocrinology & Metabolism*, 91(6), pp. 2027–2037. doi: 10.1210/jc.2005-2639.

Almine, J. F., Wise, S. G. and Weiss, A. S. (2012) 'Elastin signaling in wound repair', *Birth Defects Research Part C: Embryo Today: Reviews*, 96(3), pp. 248–257. doi: 10.1002/bdrc.21016.

Altenberg, L. (2015) 'Statistical Problems in a Paper on Variation In Cancer Risk Among Tissues, and New Discoveries', (2015). Available at: http://arxiv.org/abs/1501.04605.

Althaus, K. and Greinacher, A. (2009) 'MYH9-Related Platelet Disorders', *Seminars in Thrombosis and Hemostasis*, 35(02), pp. 189–203. doi: 10.1055/s-0029-1220327.

Andreozzi, M. *et al.* (2014) 'VEGFA gene locus analysis across 80 human tumour types reveals gene amplification in several neoplastic entities', *Angiogenesis*, 17(3), pp. 519–527. doi: 10.1007/s10456-013-9396-z.

Antoniou, A. C. *et al.* (2008) 'The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions', *British Journal of Cancer*, 98(8), pp. 1457–1466. doi: 10.1038/sj.bjc.6604305.

Antoniou, A. C. *et al.* (2014) 'Breast-Cancer Risk in Families with Mutations in PALB2', *New England Journal of Medicine*, 371(6), pp. 497–506. doi: 10.1056/NEJMoa1400382.

Ashford, N. A. *et al.* (2015) 'Cancer risk: role of environment.', *Science (New York, N.Y.)*, 347(6223), p. 727. doi: 10.1126/science.aaa6246.

Ashinoff, R., Jacobson, M. and Belsito, D. V. (1993) 'Rombo syndrome: A second case report and review', *Journal of the American Academy of Dermatology*. American Academy of Dermatology, Inc., 28(6), pp. 1011–1014. doi: 10.1016/S0190-9622(08)80656-1.

Asiaf, A. *et al.* (2014) 'Loss of expression and aberrant methylation of the CDH1 (E-cadherin) gene in breast cancer patients from Kashmir', *Asian Pacific Journal of Cancer Prevention*, 15(15), pp. 6397–6403. doi: 10.7314/APJCP.2014.15.15.6397.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Auwera, G. A. Van Der *et al.* (2014) *From FastQ data to high confidence varant calls: the Genome Analysis Toolkit best practices pipeline, Curr Protoc Bioinformatics*. doi: 10.1002/0471250953.bi1110s43.From.

Ayala-Ramirez, M. *et al.* (2013) 'Adrenocortical carcinoma: Clinical outcomes and prognosis of 330 patients at a tertiary care Center', *European Journal of Endocrinology*, 169(6), pp. 891–899. doi: 10.1530/EJE-13-0519.

Barrett, T. *et al.* (2009) 'NCBI GEO: Archive for high-throughput functional genomic data', *Nucleic Acids Research*, 37(SUPPL. 1), pp. 885–890. doi: 10.1093/nar/gkn764.

Bass, A. J. *et al.* (2014) 'Comprehensive molecular characterization of gastric adenocarcinoma', *Nature*. Nature Publishing Group, 513(7517), pp. 202–209. doi: 10.1038/nature13480.

Belmont, J. W. *et al.* (2005) 'A haplotype map of the human genome', *Nature*, 437(7063), pp. 1299–1320. doi: 10.1038/nature04226.

Bergenstal, D. M. *et al.* (1960) 'CHEMOTHERAPY OF ADRENOCORTICAL CANCER WITH o,p′DDD', *Annals of Internal Medicine*, 53(4), p. 672. doi: 10.7326/0003-4819-53-4-672.

Bernstein, J. L. *et al.* (1992) 'Risk Factors Predicting the Incidence of Second Primary Breast Cancer among Women Diagnosed with a First Primary Breast Cancer', *American Journal of Epidemiology*, 136(8), pp. 925–936.

Bernstein, J. L. *et al.* (2004) 'Study design: evaluating gene-environment interactions in the etiology of breast cancer - the WECARE study.', *Breast cancer research : BCR*, 6(3), pp. R199-214. doi: 10.1186/bcr771.

Bernstein, J. L. *et al.* (2010) 'Radiation Exposure, the ATM Gene, and Contralateral Breast Cancer in the Women's Environmental Cancer and Radiation Epidemiology Study', *JNCI Journal of the National Cancer Institute*, 102(7), pp. 475–483. doi: 10.1093/jnci/djq055.

Berruti, A. *et al.* (2005) 'Etoposide, doxorubicin and cisplatin plus mitotane in the treatment of advanced adrenocortical carcinoma: A large prospective phase II trial', *Endocrine-Related Cancer*, 12(3), pp. 657–666. doi: 10.1677/erc.1.01025.

Betapudi, V. (2014) 'Life without double-headed non-muscle myosin II motor proteins.', *Frontiers in chemistry*, 2(July), p. 45. doi: 10.3389/fchem.2014.00045.

Betts, M. J. *et al.* (2015) 'Mechismo: Predicting the mechanistic impact of mutations and modifications on molecular interactions', *Nucleic Acids Research*, 43(2), p. e10. doi: 10.1093/nar/gku1094.

Bharwani, N. *et al.* (2011) 'Adrenocortical carcinoma: The range of appearances on CT and MRI', *American Journal of Roentgenology*, 196(6), pp. 706–714. doi: 10.2214/AJR.10.5540.

bhushann Meka, P. *et al.* (2016) 'Influence of BCL2-938 C>A promoter polymorphism and BCL2 gene expression on the progression of breast cancer', *Tumor Biology*. Tumor Biology, 37(5), pp. 6905–6912. doi: 10.1007/s13277-015-4554-0.

Bilimoria, K. Y. *et al.* (2008) 'Adrenocortical carcinoma in the United States: Treatment utilization and prognostic factors', *Cancer*, 113(11), pp. 3130–3136. doi: 10.1002/cncr.23886.

Binns, D. *et al.* (2009) 'QuickGO: a web-based tool for Gene Ontology searching', *Bioinformatics*, 25(22), pp. 3045–3046. doi: 10.1093/bioinformatics/btp536.

Birch, J. M. *et al.* (1998) 'Cancer phenotype correlates with constitutional TP53 genotype in families with the Li-Fraumeni syndrome', *Oncogene*, 17(9), pp. 1061–1068. doi: 10.1038/sj.onc.1202033.

Bisceglia, M. *et al.* (2004) 'Adrenocortical oncocytic tumors: report of 10 cases and review of the literature.', *International journal of surgical pathology*, 12(3), pp. 231–43. doi: 10.1177/106689690401200304.

Blake, J. A. *et al.* (2015) 'Gene ontology consortium: Going forward', *Nucleic Acids Research*, 43(D1), pp. D1049–D1056. doi: 10.1093/nar/gku1179.

Blake, J. A. *et al.* (2017) 'Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse', *Nucleic Acids Research*, 45(D1), pp. D723–D729. doi: 10.1093/nar/gkw1040.

Boice Jr., J. D. *et al.* (1991) 'Frequent Chest X-Ray Fluoroscopy and Breast Cancer Incidence among Tuberculosis Patients in Massachusetts', *Radiation Research*, 125(2), pp. 214–222. doi: 10.2307/3577890.

Boice Jr., J. D. *et al.* (1992) 'Cancer in the contralateral breast after radiotherapy for breast cancer.', *N. Engl. J. Med.*, 326(12).

Bond, J. E. *et al.* (2011) 'Temporal spatial expression and function of non-muscle myosin II isoforms IIA and IIB in scar remodeling', *Laboratory Investigation*, 91(4), pp. 499–508. doi: 10.1038/labinvest.2010.181.

Brewer, M. H. *et al.* (2014) 'Haplotype-specific modulation of a SOX10/CREB response element at the Charcot-Marie-Tooth disease type 4C locus SH3TC2', *Human molecular genetics*, 23(19), pp. 5171–5187. doi: 10.1093/hmg/ddu240.

Brown, F. M. *et al.* (2000) 'Myxoid Neoplasms of the Adrenal Cortex', *The American Journal of Surgical Pathology*, 24(3), pp. 396–401. doi: 10.1097/00000478-200003000-00008.

Brown, K. R. and Jurisica, I. (2005) 'Online predicted human interaction database', *Bioinformatics*, 21(9), pp. 2076–2082. doi: 10.1093/bioinformatics/bti273.

Bubien, V. *et al.* (2013) 'High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome', *Journal of Medical Genetics*, 50(4), pp. 255–263. doi: 10.1136/jmedgenet-2012-101339.

Burnichon, N. *et al.* (2010) 'SDHA is a tumor suppressor gene causing paraganglioma', *Human Molecular Genetics*, 19(15), pp. 3011–3020. doi: 10.1093/hmg/ddq206.

Cai, L. *et al.* (2016) 'In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data', *Scientific Reports*. Nature Publishing Group, 6(November), pp. 1–9. doi: 10.1038/srep36540.

Caldas, C. *et al.* (1999) 'Familial gastric cancer: overview and guidelines for management.', *Journal of medical genetics*, 36(12), pp. 873–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1734270&tool=pmcentrez&rendertype=abstract.

Van Camp, J. K. *et al.* (2014) 'Wnt Signaling and the Control of Human Stem Cell Fate', *Stem Cell Reviews and Reports*, 10(2), pp. 207–229. doi: 10.1007/s12015-013-9486-8.

Cancer Research UK (2016) *Cancer Research UK Mortality Statistics*. Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared.

Cannon, J. R. (1981) 'Eruptive Syringomas', *Arch Dermatol*, 117, pp. 2–3.

Cardoso, J. C. and Calonje, E. (2015) 'Malignant sweat gland tumours: an update', *Histopathology*, 67(5), pp. 589–606. doi: 10.1111/his.12767.

Carmi, S. *et al.* (2014) 'Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins', *Nature Communications*, 5, pp. 1–9. doi: 10.1038/ncomms5835.

Carroll, J. C. *et al.* (2008) 'Hereditary breast and ovarian cancers.', *Canadian family physician Médecin de famille canadien*, 54(12), pp. 1691–2. doi: 10.1007/s00292-010-1355-5.

Carson, A. R. *et al.* (2014) 'Effective filtering strategies to improve data quality from population-based whole exome sequencing studies', *BMC Nephrology*, 15(1), pp. 1–15. doi: 10.1186/1471-2105-15-125.

Chae, Y. K. *et al.* (2016) 'Genomic landscape of DNA repair genes in cancer', 7(17).

Chajès, V., Jenab, M. and Romieu, I. (2011) 'Plasma phospholipid fatty acid concentrations and risk of gastric adenocarcinomas in the European Prospective Investigation into Cancer and Nutrition (EPIC-', *Am J clin Nutr*, (1), pp. 1304–13. doi: 10.3945/ajcn.110.005892.Study.

Chatr-Aryamontri, A. *et al.* (2015) 'The BioGRID interaction database: 2015 update', *Nucleic Acids Research*, 43(D1), pp. D470–D478. doi: 10.1093/nar/gku1204.

Chen, K. *et al.* (2009) 'BreakDancer: An algorithm for high-resolution mapping of genomic structural variation', *Nature Methods*. Nature Publishing Group, 6(9), pp. 677–681. doi: 10.1038/nmeth.1363.

Chiller, K. *et al.* (2000) 'Microcystic Adnexal Carcinoma: Forty-eight Cases, Their Treatment, and Their Outcome', *Arch Dermatol*, 136(11), pp. 1355–1359. doi: 10.1097/GOX.0000000000000195.

Cho, S. Y. *et al.* (2017) 'Sporadic Early-Onset Diffuse Gastric Cancers Have High Frequency of Somatic CDH1 Alterations, but Low Frequency of Somatic RHOA Mutations Compared With Late-Onset Cancers', *Gastroenterology*, 153(2), p. 536–549.e26. doi: 10.1053/j.gastro.2017.05.012.

Chun, Y. S. *et al.* (2001) 'Germline E-cadherin gene mutations: Is prophylactic total gastrectomy indicated?', *Cancer*, 92(1), pp. 181–187. doi: 10.1002/1097-0142(20010701)92:1<181::AID-CNCR1307>3.0.CO;2-J.

Chung, S. *et al.* (2009) 'Overexpressing PKIB in prostate cancer promotes its aggressiveness by linking between PKA and Akt pathways', *Oncogene*. Nature Publishing Group, 28(32), pp. 2849–2859. doi: 10.1038/onc.2009.144.

Cibulskis, K. *et al.* (2013) 'Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples', *Nature Biotechnology*. Nature Publishing Group, 31(3), pp. 213–219. doi: 10.1038/nbt.2514.

Clark, K. *et al.* (2008) 'TRPM7 Regulates Myosin IIA Filament Stability and Protein Localization by Heavy Chain Phosphorylation', *Journal of Molecular Biology*, 378(4), pp. 790–803. doi: 10.1016/j.jmb.2008.02.057.

Coli, A. *et al.* (2010) 'Sarcomatoid carcinoma of the adrenal gland: A case report and review of literature', *Pathology Research and Practice*. Elsevier, 206(1), pp. 59–65. doi: 10.1016/j.prp.2009.02.012.

Crowson, a N., Magro, C. M. and Mihm, M. C. (2006) 'Malignant adnexal neoplasms', *Modern Pathology*, 19, pp. S93–S126. doi: 10.1038/modpathol.3800511.

Cybulski, C., Nazarali, S. and Narod, S. a. (2014) 'Multiple primary cancers as a guide to heritability', *International Journal of Cancer*, 135(8), pp. 1756–1763. doi: 10.1002/ijc.28988.

Dang, V. T. *et al.* (2008) 'Identification of human haploinsufficient genes and their genomic proximity to segmental duplications', *European Journal of Human Genetics*, 16(11), pp. 1350–1357. doi: 10.1038/ejhg.2008.111.

Datta, A. *et al.* (2014) 'SPHK1 regulates proliferation and survival responses in triple-negative breast cancer', *Oncotarget*, 5(15), pp. 5920–5933. doi: 10.18632/oncotarget.1874.

Davies, M. A. and Samuels, Y. (2010) 'Analysis of the genome to personalize therapy for melanoma.', *Oncogene*. Nature Publishing Group, 29(41), pp. 5545–55. doi: 10.1038/onc.2010.323.

Easton, D. F. (1999) 'How many more breast cancer predisposition genes are there?', *Breast cancer research : BCR*, 1(1), pp. 14–7. doi: 10.1186/bcr6.

Easton, D. F. *et al.* (2007) 'Genome-wide association study identifies novel breast cancer susceptibility loci', *Nature*, 447(7148), pp. 1087–1093. doi: 10.1038/nature05887.

Easton, D. F. *et al.* (2015) 'Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk', *The New England journal of medicine*, 372(23), pp. 1–15. doi: 10.1056/NEJMsr1501341.

Economopoulou, P., Dimitriadis, G. and Psyrri, a. (2015) 'Beyond BRCA: New hereditary breast cancer susceptibility genes', *Cancer Treatment Reviews*. Elsevier Ltd, 41(1), pp. 1–8. doi: 10.1016/j.ctrv.2014.10.008.

el-Naggar, A. K., Evans, D. B. and Mackay, B. (1991) 'Oncocytic adrenal cortical carcinoma.', *Ultrastructural pathology*, 15(4–5), pp. 549–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1755111.

Elmi, M. *et al.* (2018) 'Concurrent risk-reduction surgery in patients with increased lifetime risk for breast and ovarian cancer: an analysis of the National Surgical Quality Improvement Program (NSQIP) database.', *Breast cancer research and treatment*. Springer US, 0(0), p. 0. doi: 10.1007/s10549-018-4818-7.

Else, T. *et al.* (2014) 'Adrenocortical carcinoma', *Endocrine Reviews*, 35(2), pp. 282–326. doi: 10.1210/er.2013-1029.

Engel, C. *et al.* (2012) 'Risks of less common cancers in proven mutation carriers with lynch syndrome', *Journal of Clinical Oncology*, 30(35), pp. 4409–4415. doi: 10.1200/JCO.2012.43.2278.

Fassnacht, M., Kroiss, M. and Allolio, B. (2013) 'Update in adrenocortical carcinoma', *Journal of Clinical Endocrinology and Metabolism*, 98(12), pp. 4551–4564. doi: 10.1210/jc.2013-3020.

Fiers, W. *et al.* (1976) 'Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene', *Nature*, 260(5551), pp. 500–507. doi: 10.1038/260500a0.

Fitzgerald, R. C. *et al.* (2010) 'Hereditary diffuse gastric cancer: Updated consensus guidelines for clinical management and directions for future research', *Journal of Medical Genetics*, 47(7), pp. 436–444. doi: 10.1136/jmg.2009.074237.

Forbes, S. A. *et al.* (2017) 'COSMIC: Somatic cancer genetics at high-resolution', *Nucleic Acids Research*, 45(D1), pp. D777–D783. doi: 10.1093/nar/gkw1121.

Fromer, M. *et al.* (2012) 'Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth', *American Journal of Human Genetics*. The American Society of Human Genetics, 91(4), pp. 597–607. doi: 10.1016/j.ajhg.2012.08.005.

Fromer, M. and Purcell, S. M. (2014) 'Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data', in *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 7.23.1-7.23.21. doi: 10.1002/0471142905.hg0723s81.

Gabillot-Carre, M. *et al.* (2006) 'Microcystic Adnexal Carcinoma: Report of Seven Cases Including One with Lung Metastasis', *Dermatology*, 212(3), pp. 221–228. doi: 10.1159/000091248.

Gemenetzidis, E. *et al.* (2010) 'Induction of Human Epithelial Stem/Progenitor Expansion by FOXM1', *Cancer Research*, 70(22), pp. 9515–9526. doi: 10.1158/0008-5472.CAN-10-2173.

Giordano, T. J. *et al.* (2003) 'Distinct transcriptional profiles of adrenocortical tumors uncovered by DNA microarray analysis', *American Journal of Pathology*, 162(2), pp. 521–531. doi: 10.1016/S0002-9440(10)63846-1.

Gnarra, J. R. *et al.* (1996) 'Post-transcriptional regulation of vascular endothelial growth factor mRNA by the product of the VHL tumor suppressor gene.', *Proceedings of the National Academy of Sciences of the United States of America*, 93(20), pp. 10589–10594. doi: 10.1073/pnas.93.20.10589.

Goldstein, D. J., Barr, R. J. and Santa Cruz, D. J. (1982) 'Microcystic adnexal carcinoma: a distinct clinicopathologic entity.', *Cancer*, 50(3), pp. 566–572. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7093897.

González, C. A. *et al.* (2003) 'Smoking and the risk of gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC)', *International Journal of Cancer*, 107(4), pp. 629–634. doi: 10.1002/ijc.11426.

Gonzalez, K. D. *et al.* (2009) 'Beyond li fraumeni syndrome: Clinical characteristics of families with p53 germline mutations', *Journal of Clinical Oncology*, 27(8), pp. 1250–1256. doi: 10.1200/JCO.2008.16.6959.

Gonzalez, R. J. *et al.* (2007) 'Response to mitotane predicts outcome in patients with recurrent adrenal cortical carcinoma', *Surgery*, 142(6), pp. 867–875. doi: 10.1016/j.surg.2007.09.006.

Gossage, L., Eisen, T. and Maher, E. R. (2015) 'VHL, the story of a tumour suppressor gene', *Nat Rev Cancer*. Nature Publishing Group, 15(1), pp. 55–64. doi: 10.1038/nrc3844.

Gotay, C., Dummer, T. and Spinelli, J. (2015) 'Cancer risk: prevention is crucial.', *Science (New York, N.Y.)*, 347(6223), p. 728. doi: 10.1126/science.aaa6462.

Green, M. *et al.* (2014) 'Microcystic adnexal carcinoma in the axilla of an 18-year-old woman', *Pediatric Dermatology*, 31(6), pp. e145–e148. doi: 10.1111/pde.12430.

Gudbjartsson, D. F. *et al.* (2015) 'Large-scale whole-genome sequencing of the Icelandic population', *Nature Genetics*. Nature Publishing Group, 47(5), pp. 435–444. doi: 10.1038/ng.3247.

Guilford, P. *et al.* (1998) 'E-cadherin germline mutations in familial gastric cancer.', *Nature*, 392(6674), pp. 402–405. doi: 10.1038/32918.

Guilford, P. *et al.* (2007) 'A short guide to hereditary diffuse gastric cancer.', *Hereditary cancer in clinical practice*, 5(4), pp. 183–94. doi: 10.1186/1897-4287-5-4-183.

Hanahan, D. and Weinberg, R. A. (2000) 'The hallmarks of cancer.', *Cell*, 100(1), pp. 57–70. doi: 10.1007/s00262-010-0968-0.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*. Elsevier Inc., 144(5), pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hansford, S. *et al.* (2015) 'Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond.', *JAMA oncology*, 1(1), pp. 23–32. doi: 10.1001/jamaoncol.2014.168.

Harker, M. (2013) 'Psychological sweating: A systematic review focused on aetiology and cutaneous response', *Skin Pharmacology and Physiology*, 26(2), pp. 92–100. doi: 10.1159/000346930.

Hartman, M. *et al.* (2005) 'Genetic implications of bilateral breast cancer: a population based cohort study.', *The lancet oncology*, 6(6), pp. 377–82. doi: 10.1016/S1470-2045(05)70174-1.

Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*. The Authors, 107(1), pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Heemskerk-Gerritsen, B. A. M. *et al.* (2013) 'Substantial breast cancer risk reduction and potential survival benefit after bilateral mastectomy when compared with surveillance in healthy BRCA1 and BRCA2 mutation carriers: A prospective analysis', *Annals of Oncology*, 24(8), pp. 2029–2035. doi: 10.1093/annonc/mdt134.

Helgason, H. *et al.* (2015) 'Loss-of-function variants in ATM confer risk of gastric cancer', *Nature Genetics*. Nature Publishing Group, 47(8), pp. 906–910. doi: 10.1038/ng.3342.

Hemminki, A. *et al.* (1998) 'A serine/threonine kinase gene defective in Peutz–Jeghers syndrome', *Nature*, 391(6663), pp. 184–187. doi: 10.1038/34432.

Heppner, C. *et al.* (1999) 'MEN1 gene analysis in sporadic adrenocortical neoplasms', *Journal of Clinical Endocrinology and Metabolism*, 84(1), pp. 216–219. doi: 10.1210/jc.84.1.216.

Herrmann, L. J. M. *et al.* (2012) 'TP53 germline mutations in adult patients with adrenocortical carcinoma', *Journal of Clinical Endocrinology and Metabolism*, 97(3), pp. 476–485. doi: 10.1210/jc.2011-1982.

Hibbs, R. (1958) 'The fine structure of human eccrine sweat glands', *American Journal of Anatomy*.

doi: 10.1002/aja.1001030204.

Higgs, J. E. *et al.* (2015) 'The BRCA2 polymorphic stop codon: stuff or nonsense?', *Journal of medical genetics*, 52(9), pp. 642–5. doi: 10.1136/jmedgenet-2015-103206.

Hoang, M. P., Ayala, A. G. and Albores-Saavedra, J. (2002) 'Oncocytic adrenocortical carcinoma: a morphologic, immunohistochemical and ultrastructural study of four cases.', *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 15(9), pp. 973–8. doi: 10.1038/modpathol.3880638.

Hocker, T. and Tsao, H. (2007) 'Ultraviolet radiation and melanoma: a systematic review and analysis of reported sequence variants.', *Human mutation*, 28(6), pp. 578–88. doi: 10.1002/humu.20481.

Hoffmann, T. J., Marini, N. J. and Witte, J. S. (2010) 'Comprehensive approach to analyzing rare genetic variants', *PLoS ONE*, 5(11). doi: 10.1371/journal.pone.0013584.

Holley, R. W. *et al.* (1965) 'Structure of a Ribonucleic Acid', *Science*, 147(3664), pp. 1462–1465. doi: 10.1126/science.147.3664.1462.

Horvath, A. *et al.* (2006) 'A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11A4 (PDE11A) in individuals with adrenocortical hyperplasia', *Nature Genetics*, 38(7), pp. 794–800. doi: 10.1038/ng1809.

Hu, A., Wang, F. and Sellers, J. R. (2002) 'Mutations in human nonmuscle myosin IIA found in patients with May-Hegglin anomaly and Fechtner syndrome result in impaired enzymatic function', *Journal of Biological Chemistry*, 277(48), pp. 46512–46517. doi: 10.1074/jbc.M208506200.

Huang, K. *et al.* (2018) 'Pathogenic Germline Variants in 10,389 Adult Cancers', *Cell*, pp. 1–16.

Hunt, R. H. *et al.* (2011) 'Helicobacter pylori in developing countries. World gastroenterology organisation global guideline', *Journal of Gastrointestinal and Liver Diseases*, 20(3), pp. 299–304. doi: 10.1097/MCG.0b013e31820fb8f6.

Huo, D. *et al.* (2009) 'Prediction of BRCA mutations using the BRCAPRO model in clinic-based African American, hispanic, and other minority families in the United States', *Journal of Clinical Oncology*, 27(8), pp. 1184–1190. doi: 10.1200/JCO.2008.17.5869.

Icard, P. *et al.* (1992) 'Adrenocortical carcinoma in surgically treated patients: a retrospective study on 156 cases by the French Association of Endocrine Surgery.', *Surgery*, 112(6), pp. 972-9; discussion 979-80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1455322.

Inskip, M. and Magee, J. (2015) 'Microcystic adnexal carcinoma of the cheek—a case report with dermatoscopy and dermatopathology', *Dermatology Practical & Conceptual*, 5(1), pp. 43–46. doi: 10.5826/dpc.0501a07.

Ishiguro, H. *et al.* (2016) 'Decreased expression of CDH1 or CTNNB1 affects poor prognosis of patients with esophageal cancer', *World Journal of Surgical Oncology*. World Journal of Surgical Oncology, 14(1), pp. 1–8. doi: 10.1186/s12957-016-0956-8.

Jackson, S. P. and Bartek, J. (2009) 'The DNA-damage response in human biology and disease', *Nature*. Nature Publishing Group, 461(7267), pp. 1071–1078. doi: 10.1038/nature08467.

Ji, S.-G. *et al.* (2016) 'Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease', *Nature Genetics*. Nature Publishing Group, 49(2), pp. 269–273. doi: 10.1038/ng.3745.

Jia, R. *et al.* (2010) 'SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance', *International Journal of Biological Sciences*, 6(7), pp. 806–826. doi: 10.7150/ijbs.6.806.

Jou, W. M. *et al.* (1972) 'Nucleotide sequence of the gene coding for the bacteriophage MS2 coat

protein', *Nature*, 237(5350), pp. 82–88. doi: 10.1038/237082a0.

de Joussineau, C. *et al.* (2012) 'The cAMP pathway and the control of adrenocortical development and growth', *Molecular and Cellular Endocrinology*, 351(1), pp. 28–36. doi: 10.1016/j.mce.2011.10.006.

Kenny, E. E. *et al.* (2012) 'A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci', *PLoS Genetics*, 8(3). doi: 10.1371/journal.pgen.1002559.

Kets, C. M. *et al.* (2009) 'Compound heterozygosity for two MSH2 mutations suggests mild consequences of the initiation codon variant c.1A>G of MSH2', *European Journal of Human Genetics*, 17(2), pp. 159–164. doi: 10.1038/ejhg.2008.153.

Kim, D. *et al.* (2013) 'TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*. BioMed Central Ltd, 14(4), p. R36. doi: 10.1186/gb-2013-14-4-r36.

Klein, C. A. (2009) 'Parallel progression of primary tumours and metastases.', *Nature reviews. Cancer*, 9(4), pp. 302–12. doi: 10.1038/nrc2627.

Koboldt, D. C. *et al.* (2010) 'Challenges of sequencing human genomes', *Briefings in Bioinformatics*, 11(5), pp. 484–498. doi: 10.1093/bib/bbq016.

Kollias, J. *et al.* (2001) 'Prognostic significance of synchronous and metachronous bilateral breast cancer.', *World Journal of Surgery*, 25(9), pp. 1117–1124. doi: 10.1007/s00268-001-0091-7.

Kreimann, E. L. *et al.* (2015) 'A novel splicing mutation in the SLC9A3R1 gene in tumors from ovarian cancer patients', *Oncology Letters*, 10(6), pp. 3722–3726. doi: 10.3892/ol.2015.3796.

Kriajevska, M. *et al.* (2000) 'Metastasis-associated protein Mts1 (S100A4) inhibits CK2-mediated phosphorylation and self-assembly of the heavy chain of nonmuscle myosin', *Biochimica et Biophysica Acta - Molecular Cell Research*, 1498(2–3), pp. 252–263. doi: 10.1016/S0167-4889(00)00100-2.

De Krijger, R. R. and Papathomas, T. G. (2012) 'Adrenocortical neoplasia: Evolving concepts in tumorigenesis with an emphasis on adrenal cortical carcinoma variants', *Virchows Archiv*, 460(1), pp. 9–18. doi: 10.1007/s00428-011-1166-y.

Kuchenbaecker, K. B. *et al.* (2017) 'Risks of Breast, Ovarian, and Contralateral Breast Cancer for *BRCA1* and *BRCA2* Mutation Carriers', *Jama*, 317(23), p. 2402. doi: 10.1001/jama.2017.7112.

Kumar, P., Henikoff, S. and Ng, P. C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nature Protocols*, 4(7), pp. 1073–1082. doi: 10.1038/nprot.2009.86.

Kurek, R. *et al.* (2001) 'Local recurrence of an oncocytic adrenocortical carcinoma with ovary metastasis.', *The Journal of urology*, 166(3), p. 985. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11490264.

Kurian, A. W. *et al.* (2009) 'Second primary breast cancer occurrence according to hormone receptor status', *Journal of the National Cancer Institute*, 101(15), pp. 1058–1065. doi: 10.1093/jnci/djp181.

Lalloo, F. and Evans, D. G. (2012) 'Familial Breast Cancer', *Clinical Genetics*, 82(2), pp. 105–114. doi: 10.1111/j.1399-0004.2012.01859.x.

Lan, Q. *et al.* (2012) 'Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia.', *Nature genetics*. Nature Publishing Group, 44(12), pp. 1330–5. doi: 10.1038/ng.2456.

Landrum, M. J. *et al.* (2018) 'ClinVar: Improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D1062–D1067. doi:

10.1093/nar/gkx1153.

Langlands, F. *et al.* (2016) 'Contralateral breast cancer: Incidence according to ductal or lobular phenotype of the primary', *Clinical Radiology*. The Royal College of Radiologists, 71(2), pp. 159–163. doi: 10.1016/j.crad.2015.10.030.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lassmann, T., Hayashizaki, Y. and Daub, C. O. (2011) 'SAMStat: Monitoring biases in next generation sequencing data', *Bioinformatics*, 27(1), pp. 130–131. doi: 10.1093/bioinformatics/btq614.

LaTorre, G. *et al.* (2009) 'Smoking status and gastric cancer risk:an update meta-analysis of case-control studies published in the past ten years.', *Tumori*, Jan-Feb 95(1), pp. 13–22. doi: 10.1700/410.4843.

Latronico, A. C. and Chrousos, G. P. (1997) 'Extensive personal experience: adrenocortical tumors.', *The Journal of clinical endocrinology and metabolism*, 82(5), pp. 1317–24. doi: 10.1210/jcem.82.5.3921.

Lau, J. and Haber, R. M. (2013) 'Familial Eruptive Syringomas: Case Report and Review of the Literature', *Journal of Cutaneous Medicine and Surgery*, 17(2), pp. 84–88. doi: 10.2310/7750.2012.12027.

LeBoit, P. E. and Sexton, M. (1993) 'Microcystic adnexal carcinoma of the skin', *Journal of the American Academy of Dermatology*. American Academy of Dermatology, Inc., 29(4), pp. 609–618. doi: 10.1016/0190-9622(93)70228-L.

Lee, A. J. *et al.* (2016) 'Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model', *Genetics in Medicine*, 18(12), pp. 1190–1198. doi: 10.1038/gim.2016.31.

Lee, S. *et al.* (2012) 'Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies', *American Journal of Human Genetics*, 91(2), pp. 224–237. doi: 10.1016/j.ajhg.2012.06.007.

Lee, Y. Y. and Derakhshan, M. H. (2013) 'Environmental and lifestyle risk factors of gastric cancer', *Archives of Iranian Medicine*, 16(6), pp. 358–365. doi: 013166/AIM.0010.

Leegte, B. (2005) 'Phenotypic expression of double heterozygosity for BRCA1 and BRCA2 germline mutations', *Journal of Medical Genetics*, 42(3), pp. e20–e20. doi: 10.1136/jmg.2004.027243.

Leinonen, R., Sugawara, H. and Shumway, M. (2011) 'The sequence read archive', *Nucleic Acids Research*, 39(SUPPL. 1), pp. 2010–2012. doi: 10.1093/nar/gkq1019.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*. Nature Publishing Group, 536(7616), pp. 285–291. doi: 10.1038/nature19057.

Lewis, F. R. *et al.* (2001) 'Prophylactic total gastrectomy for familial gastric cancer', *Surgery*, 130(4), pp. 612–619. doi: 10.1067/msy.2001.117099.

Li, B. and Leal, S. (2008) 'Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data', *The American Journal of Human Genetics*, 83, pp. 311–321. doi: 10.1016/j.ajhg.2008.06.024.

Li, C. *et al.* (2015) 'The C228T mutation of TERT promoter frequently occurs in bladder cancer stem cells and contributes to tumorigenesis of bladder cancer.', *Oncotarget*, 6(23), pp. 19542–51. doi: 10.18632/oncotarget.4295.

Li, F. P. and Fraumeni, J. F. (1969) 'Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome?', *Annals of internal medicine*, 71(4), pp. 747–52. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/5360287.

Li, H. *et al.* (2009) 'The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup', *Bioinformatics (Oxford, England)*, 25(16), pp. 1–2. doi: 10.1093/bioinformatics/btp352.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27(21), pp. 2987–2993. doi: 10.1093/bioinformatics/btr509.

Li, H. and Durbin, R. (2010) 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics*, 26(5), pp. 589–595. doi: 10.1093/bioinformatics/btp698.

Liang, X. H. *et al.* (1999) 'Induction of autophagy and inhibition of tumorigenesis by beclin 1', *Nature*, 402(6762), pp. 672–676. doi: 10.1038/45257.

Libé, R. *et al.* (2011) 'Frequent phosphodiesterase 11A gene (PDE11A) defects in patients with carney complex (CNC) caused by PRKAR1A mutations: PDE11A may contribute to adrenal and testicular tumors in CNC as a modifier of the phenotype', *Journal of Clinical Endocrinology and Metabolism*, 96(1), pp. 208–214. doi: 10.1210/jc.2010-1704.

Libé, R. and Bertherat, J. (2005) 'Molecular genetics of adrenocortical tumours, from familial to sporadic diseases', *European Journal of Endocrinology*, 153(4), pp. 477–487. doi: 10.1530/eje.1.02004.

van Lier, M. G. F. *et al.* (2011) 'High cancer risk and increased mortality in patients with Peutz-Jeghers syndrome.', *Gut*, 60(2), pp. 141–7. doi: 10.1136/gut.2010.223750.

Lindahl, T. and Barnes, D. E. (2000) 'Repair of endogenous DNA damage.', *Cold Spring Harbor symposia on quantitative biology*, 65, pp. 127–33. doi: 10.1101/sqb.2000.65.127.

Liu, H. *et al.* (2015) 'SLC9A3R1 stimulates autophagy via BECN1 stabilization in breast cancer cells', *Autophagy*, 11(12), pp. 2323–2334. doi: 10.1080/15548627.2015.1074372.

Lizarraga, I. M. *et al.* (2013) 'Review of risk factors for the development of contralateral breast cancer', *American Journal of Surgery*. Elsevier Inc, 206(5), pp. 704–708. doi: 10.1016/j.amjsurg.2013.08.002.

Loeb, L. A. (2001) 'Perspectives in Cancer Research A Mutator Phenotype in Cancer', *Cancer Res.*, 61(8), pp. 3230–3239.

Loeb, L. A. (2016) 'Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences', *Cancer Research*, 76(8), pp. 2057–2059. doi: 10.1158/0008-5472.CAN-16-0794.

Lopes, M. C. *et al.* (2012) 'A combined functional annotation score for non-synonymous variants', *Human Heredity*, 73(1), pp. 47–51. doi: 10.1159/000334984.

Lu, C. *et al.* (2015) 'Patterns and functional implications of rare germline variants across 12 cancer types', *Nature Communications*. Nature Publishing Group, 6, p. 10086. doi: 10.1038/ncomms10086.

Lu, C. and Fuchs, E. (2014) 'Sweat gland progenitors in development, homeostasis, and wound repair', *Cold Spring Harbor Perspectives in Medicine*, 4(2), pp. 1–18. doi: 10.1101/cshperspect.a015222.

Luo, Y. *et al.* (2017) 'Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7', *Nature Genetics*. Nature Publishing Group, 49(2), pp. 186–192. doi: 10.1038/ng.3761.

Luton, J. P. *et al.* (1990) 'Clinical features of adrenocortical carcinoma, prognostic factors, and the effect of mitotane therapy.', *The New England journal of medicine*, 322(17), pp. 1195–201. doi: 10.1056/NEJM199004263221705.

Lynch, H. T. *et al.* (1966) 'Hereditary Factors in Cancer', *Archives of Internal Medicine*, 117(2), p. 206. doi: 10.1001/archinte.1966.03870080050009.

Lynch, H. T. *et al.* (1977) 'Familial cancer syndromes':, *Cancer*, 39, pp. 1867–1881.

Lynch, H. T. *et al.* (1985) 'Pancreatic carcinoma and hereditary nonpolyposis colorectal cancer: a family study.', *British journal of cancer*, 52(2), pp. 271–3. doi: 10.1038/bjc.1985.187.

Maczis, M., Milstien, S. and Spiegel, S. (2016) 'Sphingosine-1-phosphate and estrogen signaling in breast cancer', *Advances in Biological Regulation*. Elsevier Ltd, 60, pp. 160–165. doi: 10.1016/j.jbior.2015.09.006.

Madsen, B. E. and Browning, S. R. (2009) 'A groupwise association test for rare mutations using a weighted sum statistic', *PLoS Genetics*, 5(2). doi: 10.1371/journal.pgen.1000384.

Malone, K. E. *et al.* (2010) 'Population-Based Study of the Risk of Second Primary Contralateral Breast Cancer Associated With Carrying a Mutation in BRCA1 or BRCA2', *Journal of Clinical Oncology*, 28(14), pp. 2404–2410. doi: 10.1200/JCO.2009.24.2495.

Manolio, T. A. *et al.* (2009) 'Finding the missing heritability of complex diseases', *Nature*. Nature Publishing Group, 461(7265), pp. 747–753. doi: 10.1038/nature08494.

Mar, V. J. *et al.* (2013) 'BRAF/NRAS wild-type melanomas have a high mutation load correlating with histologic and molecular signatures of UV damage', *Clinical Cancer Research*, 19(17), pp. 4589–4598. doi: 10.1158/1078-0432.CCR-13-0398.

Maria Kałużna, E. *et al.* (2015) 'Heterozygous p.I171V mutation of the NBN gene as a risk factor for lung cancer development', *Oncology Letters*, 10(5), pp. 3300–3304. doi: 10.3892/ol.2015.3715.

Marini, M. *et al.* (2006) 'Non-muscle myosin heavy chain IIA and IIB interact and co-localize in living cells: Relevance for MYH9-related disease', *International Journal of Molecular Medicine*, 17(5), pp. 729–736.

Masciari, S. *et al.* (2011) 'Gastric cancer in individuals with Li-Fraumeni syndrome', *Genetics in Medicine*, 13(7), pp. 651–657. doi: 10.1097/GIM.0b013e31821628b6.

Masso, M. and Vaisman, I. I. (2010) 'AUTO-MUTE: Web-based tools for predicting stability changes in proteins due to single amino acid replacements', *Protein Engineering, Design and Selection*, 23(8), pp. 683–687. doi: 10.1093/protein/gzq042.

Matkovich, S. J. *et al.* (2006) 'Cardiac-specific ablation of G-protein receptor kinase 2 redefines its roles in heart development and β-adrenergic signaling', *Circulation Research*, 99(9), pp. 996–1003. doi: 10.1161/01.RES.0000247932.71270.2c.

Mayer-Jochimsen, M., Fast, S. and Tintle, N. L. (2013) 'Assessing the Impact of Differential Genotyping Errors on Rare Variant Tests of Association', *PLoS ONE*. Edited by Z. Yu, 8(3), p. e56626. doi: 10.1371/journal.pone.0056626.

Mazzolini, R. *et al.* (2012) 'Brush border Myosin Ia has tumor suppressor activity in the intestine', *Proceedings of the National Academy of Sciences*, 109, pp. 1530–1535. doi: 10.1073/pnas.1108411109.

Mazzolini, R. *et al.* (2013) 'Brush border myosin Ia inactivation in gastric but not endometrial tumors', *International Journal of Cancer*, 132(8), pp. 1790–1799. doi: 10.1002/ijc.27856.

McKay, J. D. *et al.* (2017) 'Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes', *Nature Genetics*, 49(7), pp. 1126–1132. doi: 10.1038/ng.3892.

McKenna, A. *et al.* (2010) 'The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20(9), pp. 1297–1303. doi:

10.1101/gr.107524.110.

McKinley, L. H. *et al.* (2014) 'Microcystic adnexal carcinoma: review of a potential diagnostic pitfall and management.', *Cutis*, 93(3), pp. 162–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24738099.

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*. Genome Biology, 17(1), pp. 1–14. doi: 10.1186/s13059-016-0974-4.

Mellemkjær, L. *et al.* (2008) 'Risk for contralateral breast cancer among carriers of the CHEK2*1100delC mutation in the WECARE Study', *British Journal of Cancer*, 98(4), pp. 728–733. doi: 10.1038/sj.bjc.6604228.

Métayé, T. *et al.* (2008) 'Immunohistochemical detection, regulation and antiproliferative function of G-protein-coupled receptor kinase 2 in thyroid carcinomas', *Journal of Endocrinology*, 198(1), pp. 101–110. doi: 10.1677/JOE-07-0562.

Mi, E. Z. *et al.* (2017) 'Comparative study of endoscopic surveillance in hereditary diffuse gastric cancer according to CDH1 mutation status', *Gastrointestinal Endoscopy*. American Society for Gastrointestinal Endoscopy. doi: 10.1016/j.gie.2017.06.028.

Michaelsson, G., Olsson, E. and Westermark, P. (1981) 'The Rombo syndrome: A familial disorder with vermiculate atrophoderma, milia, hypotrichosis, trichoepitheliomas, basal cell carcinomas and peripheral vasodilation with cyanosis', *Acta Dermato-Venereologica*, 61(6), pp. 497–503.

Michailidou, K. *et al.* (2015) 'Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer', *Nature Genetics*, 47(4), pp. 373–80. doi: 10.1038/ng.3242.

Michailidou, K. *et al.* (2017) 'Association analysis identifies 65 new breast cancer risk loci', *Nature*, 551(7678), pp. 92–94. doi: 10.1038/nature24284.

Michalkiewicz, E. *et al.* (2004) 'Clinical and outcome characteristics of children with adrenocortical tumors: A report from the international pediatric adrenocortical tumor registry', *Journal of Clinical Oncology*, 22(5), pp. 838–845. doi: 10.1200/JCO.2004.08.085.

Miki, Y. *et al.* (1994) 'A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.', *Science (New York, N.Y.)*, 266(5182), pp. 66–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7545954.

Milne, R. L. *et al.* (2017) 'Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer', *Nature Genetics*, 49(12), pp. 1767–1778. doi: 10.1038/ng.3785.

Moisio, A. L. *et al.* (1996) 'Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer.', *American journal of human genetics*, 59(6), pp. 1243–51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1914865&tool=pmcentrez&rendertype=abstract.

Montojo, J. *et al.* (2010) 'GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop', *Bioinformatics*, 26(22), pp. 2927–2928. doi: 10.1093/bioinformatics/btq562.

Morak, M. *et al.* (2017) 'Loss of MSH2 and MSH6 due to heterozygous germline defects in MSH3 and MSH6', *Familial Cancer*. Springer Netherlands, 16(4), pp. 491–500. doi: 10.1007/s10689-017-9975-z.

Morgenthaler, S. and Thilly, W. G. (2007) 'A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)', *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1–2), pp. 28–56. doi: 10.1016/j.mrfmmm.2006.09.003.

Mostafavi, S. *et al.* (2008) 'GeneMANIA: A real-time multiple association network integration

algorithm for predicting gene function', *Genome Biology*, 9(SUPPL. 1), pp. 1–15. doi: 10.1186/gb-2008-9-s1-s4.

Mucci, L. A. *et al.* (2016) 'Familial risk and heritability of cancer among twins in nordic countries', *JAMA - Journal of the American Medical Association*, 315(1), pp. 68–76. doi: 10.1001/jama.2015.17703.

Murota, H. *et al.* (2015) 'Sweat, the driving force behind normal skin: An emerging perspective on functional biology and regulatory mechanisms', *Journal of Dermatological Science*. Japanese Society for Investigative Dermatology, 77(1), pp. 3–10. doi: 10.1016/j.jdermsci.2014.08.011.

Muzny, D. M. *et al.* (2012) 'Comprehensive molecular characterization of human colon and rectal cancer', *Nature*. Nature Publishing Group, 487(7407), pp. 330–337. doi: 10.1038/nature11252.

Nakano, T. *et al.* (2005) 'Genetic and epigenetic alterations of the candidate tumor-suppressor gene MYO18B, on chromosome arm 22q, in colorectal cancer', *Genes Chromosomes and Cancer*, 43(2), pp. 162–171. doi: 10.1002/gcc.20180.

Narod, S. a. (2014) 'Bilateral breast cancers', *Nature Reviews Clinical Oncology*. Nature Publishing Group, 11(3), pp. 157–166. doi: 10.1038/nrclinonc.2014.3.

Nelson, A. A. and Woodard, G. (1948) 'Adrenal cortical atrophy and liver damage produced in dogs by feeding 2,2-bis-(parachloro-phenyl)-1,1-dichloroethane.', *Federation proceedings*, 7(1 Pt 1), p. 277. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18916907.

Nicolau, S. and Balus, L. (1961) 'Sur un cas de genodermatose polydysplasique [On a case of polydysplastic genodermatosis]', *Annales de dermatologie et de syphiligraphie*, 88, pp. 385–96. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14479530.

Nieuwenhuis, M. H. *et al.* (2014) 'Cancer risk and genotype-phenotype correlations in PTEN hamartoma tumor syndrome', *Familial Cancer*, 13(1), pp. 57–63. doi: 10.1007/s10689-013-9674-3.

Nishioka, M. *et al.* (2002) 'MYO18B, a candidate tumor suppressor gene at chromosome 22q12.1, deleted, mutated, and methylated in human lung cancer', *Proceedings of the National Academy of Sciences*, 99(19), pp. 12269–12274. doi: 10.1073/pnas.192445899.

Nogués, L. *et al.* (2016) 'G Protein-coupled Receptor Kinase 2 (GRK2) Promotes Breast Tumorigenesis Through a HDAC6-Pin1 Axis.', *EBioMedicine*. The Authors, 13, pp. 132–145. doi: 10.1016/j.ebiom.2016.09.030.

Norton, J. A. *et al.* (2007) 'CDH1 truncating mutations in the E-cadherin gene: An indication for total gastrectomy to treat hereditary diffuse gastric cancer', *Annals of Surgery*, 245(6), pp. 873–879. doi: 10.1097/01.sla.0000254370.29893.e4.

Notta, F. *et al.* (2016) 'A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns', *Nature*. Nature Publishing Group, 538(7625), pp. 378–382. doi: 10.1038/nature19823.

Nowell, P. (1976) 'The clonal evolution of tumor cell populations', *Science*, 194(4260), pp. 23–28. doi: 10.1126/science.959840.

Nyrén, P. and Lundin, A. (1985) 'Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis', *Analytical Biochemistry*, 151(2), pp. 504–509. doi: 10.1016/0003-2697(85)90211-8.

Nyström-Lahti, M. *et al.* (1995) 'Founding mutations and Alu-mediated recombination in hereditary colon cancer', *Nature Medicine*, 1(11), pp. 1203–1206. doi: 10.1038/nm1195-1203.

O'Callaghan, M. (2015) 'Cancer risk: accuracy of literature.', *Science (New York, N.Y.)*, 347(6223), p. 729. doi: 10.1126/science.aaa6212.

O'Rawe, J. *et al.* (2013) 'Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing', *Genome Medicine*, 5(3). doi: 10.1186/gm432.

Oda, H. *et al.* (2014) 'Aicardi-goutières syndrome is caused by IFIH1 mutations', *American Journal of Human Genetics*. The American Society of Human Genetics, 95(1), pp. 121–125. doi: 10.1016/j.ajhg.2014.06.007.

Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2015) 'Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics*, p. btv566. doi: 10.1093/bioinformatics/btv566.

Oliveira, C. *et al.* (2015) 'Familial gastric cancer: genetic susceptibility, pathology, and implications for management', *The Lancet Oncology*. Elsevier Ltd, 16(2), pp. e60–e70. doi: 10.1016/S1470-2045(14)71016-2.

Olivier, M., Hollstein, M. and Hainaut, P. (2010) 'TP53 mutations in human cancers: origins, consequences, and clinical use', *Cold Spring Harb Perspect Biol*, 2(1), p. a001008. doi: 10.1101/cshperspect.a001008.

Ongenae, K. C. *et al.* (2001) 'Microcystic adnexal carcinoma: An uncommon tumor with debatable origin', *Dermatologic Surgery*, 27(11), pp. 979–984. doi: 10.1046/j.1524-4725.2001.01061.x.

Osorio, A. *et al.* (2002) 'Loss of heterozygosity analysis at the BRCA loci in tumor samples from patients with familial breast cancer', *International Journal of Cancer*, 99(2), pp. 305–309. doi: 10.1002/ijc.10337.

Ouderkirk, J. L. and Krendel, M. (2014) 'Non-muscle myosins in tumor progression, cancer cell invasion, and metastasis', *Cytoskeleton*, 71(8), pp. 447–463. doi: 10.1002/cm.21187.

Pabinger, S. *et al.* (2014) 'A survey of tools for variant analysis of next-generation genome sequencing data', *Briefings in Bioinformatics*, 15(2), pp. 256–278. doi: 10.1093/bib/bbs086.

Pauty, J. *et al.* (2014) 'Exploring the roles of PALB2 at the crossroads of DNA repair and cancer', *Biochemical Journal*, 460(3), pp. 331–342. doi: 10.1042/BJ20140208.

Pellegrino, B. *et al.* (2016) 'Triple negative status and BRCA mutations in contralateral breast cancer: a population-based study.', *Acta bio-medica : Atenei Parmensis*, 87(1), pp. 54–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27163896.

Pern, F. *et al.* (2012) 'Mutation Analysis of BRCA1, BRCA2, PALB2 and BRD7 in a Hospital-Based Series of German Patients with Triple-Negative Breast Cancer', *PLoS ONE*, 7(10), pp. 7–12. doi: 10.1371/journal.pone.0047993.

Petitjean, A. *et al.* (2007) 'Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database', *Human Mutation*, 28(6), pp. 622–629. doi: 10.1002/humu.20495.

Pharoah, P. D. P., Guilford, P. and Caldas, C. (2001) 'Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families', *Gastroenterology*, 121(6), pp. 1348–1353. doi: 10.1053/gast.2001.29611.

Plaza-Menacho, I., Mologni, L. and McDonald, N. Q. (2014) 'Mechanisms of RET signaling in cancer: Current and future implications for targeted therapy', *Cellular Signalling*. Elsevier Inc., 26(8), pp. 1743–1752. doi: 10.1016/j.cellsig.2014.03.032.

Pon, J. R. and Marra, M. A. (2015) 'Driver and Passenger Mutations in Cancer', *Annual Review of Pathology: Mechanisms of Disease*, 10(1), pp. 25–50. doi: 10.1146/annurev-pathol-012414-040312.

Ponti, G. *et al.* (2015) 'Mismatch repair genes founder mutations and cancer susceptibility in Lynch syndrome', *Clinical Genetics*, 87(6), pp. 507–516. doi: 10.1111/cge.12529.

Poplin, R. *et al.* (2017) 'Scaling accurate genetic variant discovery to tens of thousands of samples', *bioRxiv*, p. 201178. doi: 10.1101/201178.

van der Post, R. S. *et al.* (2015) 'Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers.', *Journal of medical genetics*, 52(6), pp. 361–374. doi: 10.1136/jmedgenet-2015-103094.

Potter, J. D. and Prentice, R. L. (2015) 'Cancer risk: tumors excluded.', *Science (New York, N.Y.)*, 347(6223), p. 727. doi: 10.1126/science.aaa6507.

Powers, S., Gopalakrishnan, S. and Tintle, N. (2011) 'Assessing the impact of non-differential genotyping errors on rare variant tests of association', *Human Heredity*, 72(3), pp. 153–160. doi: 10.1159/000332222.

Pritchard, J. and Cox, N. (2002) 'The allelic architecture of human disease genes: common disease -- common variant ... or not?', *Hum. Mol. Genet.*, 11(20), pp. 2417–2423. doi: 10.1093/hmg/11.20.2417.

Prowatke, I. *et al.* (2007) 'Expression analysis of imbalanced genes in prostate carcinoma using tissue microarrays', *British Journal of Cancer*, 96(1), pp. 82–88. doi: 10.1038/sj.bjc.6603490.

Public Health England (2015) *National Cancer Intelligence Network - Rare and less common cancers*.

Rabbani, B., Tekin, M. and Mahdieh, N. (2014) 'The promise of whole-exome sequencing in medical genetics.', *Journal of human genetics*. Nature Publishing Group, 59(1), pp. 5–15. doi: 10.1038/jhg.2013.114.

Rainville, I. R. and Rana, H. Q. (2014) 'Next-Generation Sequencing for Inherited Breast Cancer Risk: Counseling through the Complexity', *Current Oncology Reports*, 16(3), pp. 1–11. doi: 10.1007/s11912-013-0371-z.

Ramus, S. J. *et al.* (2015) 'Germline mutations in the BRIP1, BARD1, PALB2, and NBN genes in women with ovarian cancer', *Journal of the National Cancer Institute*, 107(11), pp. 1–8. doi: 10.1093/jnci/djv214.

Raymond, V. M., Everett, J. N., *et al.* (2013) 'Adrenocortical carcinoma is a lynch syndrome-associated cancer', *Journal of Clinical Oncology*, 31(24), pp. 3012–3018. doi: 10.1200/JCO.2012.48.0988.

Raymond, V. M., Else, T., *et al.* (2013) 'Prevalence of Germline TP53 mutations in a prospective series of unselected patients with adrenocortical carcinoma', *Journal of Clinical Endocrinology and Metabolism*, 98(1), pp. 119–125. doi: 10.1210/jc.2012-2198.

Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein mutations: Application to cancer genomics', *Nucleic Acids Research*, 39(17), pp. 37–43. doi: 10.1093/nar/gkr407.

Ribeiro, R. C. *et al.* (2001) 'An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(16), pp. 9330–5. doi: 10.1073/pnas.161479898.

Rinella, E. S. *et al.* (2013) 'Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation', *Human Genetics*, 132(5), pp. 523–536. doi: 10.1007/s00439-013-1269-4.

RStudio Inc (2013) *Easy web applications in R*. Available at: http://shiny.rstudio.com/ (Accessed: 1 February 2018).

Ruark, E. *et al.* (2015) 'The ICR1000 UK exome series: a resource of gene variation in an outbred population', *F1000Research*, 4, p. 883. doi: 10.12688/f1000research.7049.1.

Rubin, B. *et al.* (2015) 'Mitogen-activated protein kinase pathway: Genetic analysis of 95 adrenocortical tumors', *Cancer Investigation*, 33(10), pp. 526–531. doi: 10.3109/07357907.2015.1080832.

Sahasrabudhe, R. *et al.* (2017) 'Germline Mutations in PALB2, BRCA1, and RAD51C, Which Regulate DNA Recombination Repair, in Patients With Gastric Cancer', *Gastroenterology*, 152(2017), p. 983–986.e6. doi: 10.1053/j.gastro.2016.12.010.

Saltzman, B. S. *et al.* (2012) 'Estrogen receptor, progesterone receptor, and HER2-neu expression in first primary breast cancers and risk of second primary contralateral breast cancer', *Breast Cancer Research and Treatment*, 135(3), pp. 849–855. doi: 10.1007/s10549-012-2183-5.

Sanders, D. A. *et al.* (2013) 'Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells.', *Genome biology*, 14(1), p. R6. doi: 10.1186/gb-2013-14-1-r6.

Sang, Y. S. *et al.* (2004) 'Oncocytic adrenocortical carcinomas: A pathological and immunohistochemical study of four cases in comparison with conventional adrenocortical carcinomas', *Pathology International*, 54(8), pp. 603–610. doi: 10.1111/j.1440-1827.2004.01669.x.

Sanger, F., Brownlee, G. G. and Barrell, B. G. (1965) 'A two-dimensional fractionation procedure for radioactive nucleotides', *Journal of Molecular Biology*. Academic Press Inc. (London) Ltd., 13(2), pp. IN1-IN4. doi: 10.1016/S0022-2836(65)80104-8.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Saponaro, M. *et al.* (2014) 'RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress', *Cell*, 157(5), pp. 1037–1049. doi: 10.1016/j.cell.2014.03.048.

Saposnik, B. *et al.* (2014) 'Mutation spectrum and genotype-phenotype correlations in a large French cohort of MYH9-Related Disorders.', *Molecular genetics & genomic medicine*, 2(4), pp. 297–312. doi: 10.1002/mgg3.68.

Saule, C. *et al.* (2018) 'Risk of Serous Endometrial Carcinoma in Women With Pathogenic BRCA1/2 Variant After Risk-Reducing Salpingo-Oophorectomy', *JNCI: Journal of the National Cancer Institute*, 110(2), pp. 2017–2019. doi: 10.1093/jnci/djx159.

Schaapveld, M. *et al.* (2008) 'The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: A population based study in the Netherlands', *Breast Cancer Research and Treatment*, 110(1), pp. 189–197. doi: 10.1007/s10549-007-9709-2.

Schaller, J. *et al.* (2010) 'Sweat duct proliferation associated with aggregates of elastic tissue and atrophodermia vermiculata: a simulator of microcystic adnexal carcinoma. Report of two cases†', *Journal of Cutaneous Pathology*, 37(9), pp. 1002–1009. doi: 10.1111/j.1600-0560.2010.01527.x.

Schramek, D. *et al.* (2014) 'Direct in Vivo RNAi Screen Unveils Myosin IIa as a Tumor Suppressor of Squamous Cell Carcinomas', *Science*, 343(January), pp. 309–313.

Schteingart, D. E. *et al.* (2005) 'Management of patients with adrenal cancer : recommendations of an international consensus conference', 15, pp. 667–680. doi: 10.1677/erc.

Seemanová, E. *et al.* (2007) 'Cancer risk of heterozygotes with the NBN founder mutation', *Journal of the National Cancer Institute*, 99(24), pp. 1875–1880. doi: 10.1093/jnci/djm251.

Senkus, E. *et al.* (2014) 'Are synchronous and metachronous bilateral breast cancers different? An immunohistochemical analysis aimed at intrinsic tumor phenotype', *International Journal of Clinical and Experimental Pathology*, 7(1), pp. 353–363.

Seo, I. S., Henley, J. D. and Min, K.-W. (2002) 'Peculiar cytoplasmic inclusions in oncocytic adrenal cortical tumors: an electron microscopic observation.', *Ultrastructural pathology*, 26(4), pp. 229–35. doi: 10.1080/01913120290104485.

Sereno, M. *et al.* (2011) 'Gastric tumours in hereditary cancer syndromes: Clinical features, molecular biology and strategies for prevention', *Clinical and Translational Oncology*, 13(9), pp. 599–610. doi: 10.1007/s12094-011-0705-y.

Shain, A. H. *et al.* (2015) 'Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway', *Nature Genetics*. Nature Publishing Group, 47(10), pp. 1194–1199. doi: 10.1038/ng.3382.

Sham, P. C. and Purcell, S. M. (2014) 'Statistical power and significance testing in large-scale genetic studies', *Nature Reviews Genetics*. Nature Publishing Group, 15(5), pp. 335–346. doi: 10.1038/nrg3706.

Shlien, A. *et al.* (2015) 'Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers', *Nature Genetics*. Nature Publishing Group, 47(3), pp. 257–262. doi: 10.1038/ng.3202.

Shyr, C. *et al.* (2014) 'FLAGS, frequently mutated genes in public exomes.', *BMC medical genomics*, 7, p. 64. doi: 10.1186/s12920-014-0064-y.

Siegel, R., Naishadham, D. and Jemal, A. (2013) 'Cancer statistics, 2013', *CA: A Cancer Journal for Clinicians*, 63(1), pp. 11–30. doi: 10.3322/caac.21166.

Simon Anders and Wolfgang Huber (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.

Simon, R. and Zhang, X. (2008) 'On the dynamics of breast tumor development in women carrying germline BRCA1 and BRCA2 mutations', *International Journal of Cancer*, 122(8), pp. 1916–1917. doi: 10.1002/ijc.23323.

Sisti, J. S. *et al.* (2015) 'Reproductive factors, tumor estrogen receptor status and contralateral breast cancer risk: results from the WECARE study.', *SpringerPlus*. Springer International Publishing, 4, p. 825. doi: 10.1186/s40064-015-1642-y.

Skogseid, B. *et al.* (1995) 'Adrenal lesion in multiple endocrine neoplasia type 1', *Surgery*, 118(6), pp. 1077–1082. doi: 10.1016/S0039-6060(05)80117-5.

Skol, A. D., Sasaki, M. M. and Onel, K. (2016) 'The genetics of breast cancer risk in the post-genome era: Thoughts on study design to move past BRCA and towards clinical relevance', *Breast Cancer Research*. Breast Cancer Research, 18(1), pp. 1–8. doi: 10.1186/s13058-016-0759-4.

Slavin, T. *et al.* (2017) 'Genetic Gastric Cancer Susceptibility in the International Clinical Cancer Genomics Community Research Network', *Cancer Genetics*. Elsevier Inc., 216–217, pp. 111–119. doi: 10.1016/j.cancergen.2017.08.001.

Smith, K. J. *et al.* (2001) 'Microcystic adnexal carcinoma: an immunohistochemical study including markers of proliferation and apoptosis.', *The American journal of surgical pathology*, 25(4), pp. 464–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11257620.

Song, M. and Giovannucci, E. L. (2015) 'Cancer risk: many factors contribute.', *Science (New York, N.Y.)*, 347(6223), pp. 728–9. doi: 10.1126/science.aaa6094.

Staff, S. *et al.* (2000) 'Multiple copies of mutant BRCA1 and BRCA2 alleles in breast tumors from germ-line mutation carriers.', *Genes, chromosomes & cancer*, 28(4), pp. 432–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10862052.

Van Steensel, M. A. M., Jaspers, N. G. J. and Steijlen, P. M. (2001) 'A case of Rombo syndrome', *British Journal of Dermatology*, 144(6), pp. 1215–1218. doi: 10.1046/j.1365-2133.2001.04235.x.

Stovall, M. *et al.* (2008) 'Dose to the Contralateral Breast from Radiation Therapy and Risk of Second Primary Breast Cancer in the WECARE Study', *Int J Radiat Oncol Biol Phys*, 72(4), pp. 1021–1030. doi: 10.1038/nmeth.2250.Digestion.

Stratton, M., Campbell, P. and Futreal, P. (2009) 'The cancer genome', *Nature*, 458(7239), pp. 719–724. doi: 10.1038/nature07943.The.

Stroop, S. D. and Beavo, J. A. (1991) 'Structure and function studies of the cGMP-stimulated phosphodiesterase', *Journal of Biological Chemistry*, 266(35), pp. 23802–23809.

Sudmant, P. H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81. doi: 10.1038/nature15394.

Sun, P. *et al.* (2015) 'Genetic variation in the 3'-untranslated region of NBN gene is associated with gastric cancer risk in a Chinese population', *PLoS ONE*, 10(9), pp. 1–10. doi: 10.1371/journal.pone.0139059.

Sun, Y. *et al.* (2015) 'Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome?', *Human Mutation*, 36(6), pp. 648–655. doi: 10.1002/humu.22783.

Szulkin, R. *et al.* (2015) 'Prediction of individual genetic risk to prostate cancer using a polygenic score', *Prostate*, 75(13), pp. 1467–1474. doi: 10.1002/pros.23037.

Tan, M.-H. *et al.* (2012) 'Lifetime Cancer Risks in Individuals with Germline PTEN Mutations', *Clinical Cancer Research*, 18(2), pp. 400–407. doi: 10.1158/1078-0432.CCR-11-2283.

Tanaka, K. *et al.* (2004) 'Oncocytic adrenocortical carcinoma', *Urology*, 64(2), pp. 376–377. doi: 10.1016/j.urology.2004.04.023.

Tang, H. and Thomas, P. D. (2016) 'Tools for predicting the functional impact of nonsynonymous genetic variation', *Genetics*, 203(2), pp. 635–647. doi: 10.1534/genetics.116.190033.

Tarpey, P. S. *et al.* (2013) 'Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma', *Nature Genetics*, 45(8), pp. 923–926. doi: 10.1038/ng.2668.

Taskén, K. and Aandahl, E. M. (2004) 'Localized Effects of cAMP Mediated by Distinct Routes of Protein Kinase A', *Physiological Reviews*, 84(1), pp. 137–167. doi: 10.1152/physrev.00021.2003.

Tattini, L., D'Aurizio, R. and Magi, A. (2015) 'Detection of Genomic Structural Variants from Next-Generation Sequencing Data', *Frontiers in Bioengineering and Biotechnology*, 3(June), pp. 1–8. doi: 10.3389/fbioe.2015.00092.

Teh, M. *et al.* (2002) 'FOXM1 Is a Downstream Target of Gli1 in Basal Cell Carcinomas', (4773), pp. 4773–4780.

Teh, M. T. *et al.* (2010) 'Upregulation of FOXM1 induces genomic instability in human epidermal keratinocytes.', *Mol Cancer*, 9, p. 45. doi: 1476-4598-9-45 [pii] 10.1186/1476-4598-9-45.

Teraoka, S. N. *et al.* (2011) 'Single nucleotide polymorphisms associated with risk for contralateral breast cancer in the Women's Environment, Cancer, and Radiation Epidemiology (WECARE) Study', *Breast Cancer Research*, 13(6), p. R114. doi: 10.1186/bcr3057.

Terzolo, M. *et al.* (2007) 'Adjuvant Mitotane Treatment for Adrenocortical Carcinoma', *New England Journal of Medicine*, 356(23), pp. 2372–2380. doi: 10.1056/NEJMoa063360.

The GTEx Consortium *et al.* (2015) 'The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.', *Science*, 348(6235), pp. 648–60. doi: 10.1126/science.1262110.

Tomasetti, C. and Vogelstein, B. (2015a) 'Cancer risk: role of environment—response.', *Science*, 347(6223), pp. 729–31. doi: 10.1126/science.aaa6592.

Tomasetti, C. and Vogelstein, B. (2015b) 'Variation in cancer risk among tissues can be explained by the number of stem cell divisions', *Science*, 347(6217), pp. 78–81. doi: 10.1126/science.1260825.

Tramacere, I. *et al.* (2012) 'A meta-analysis on alcohol drinking and gastric cancer risk', *Annals of Oncology*, 23(1), pp. 28–36. doi: 10.1093/annonc/mdr135.

Trapnell, C. *et al.* (2013) 'Differential analysis of gene regulation at transcript resolution with RNA-seq', *Nature Biotechnology*. Nature Publishing Group, 31(1), pp. 46–53. doi: 10.1038/nbt.2450.

Tubbs, A. and Nussenzweig, A. (2017) 'Endogenous DNA Damage as a Source of Genomic Instability in Cancer', *Cell*, 168(4), pp. 644–656. doi: 10.1016/j.cell.2017.01.002.

Turnbull, C., Ahmed, S. and Morrison, J. (2010) 'Genome-wide association study identifies five new breast cancer susceptibility loci', *Nature …*, 42(6), pp. 504–507. doi: 10.1038/ng.586.Genome-wide.

Tutt, A. *et al.* (2010) 'Oral Poly (ADP-ribose) Polymerase Inhibitor Olaparib in Patients with BRCA1 or BRCA2 Mutations and Advanced Breast Cancer: a Proof-of-concept Trial', *The Lancet*, 376(9737), pp. 235–44. doi: 10.1016/S0140-6736(10)60892-6.

Uhlen, M. *et al.* (2017) 'A pathology atlas of the human cancer transcriptome', *Science*, 357(6352). doi: 10.1126/science.aan2507.

Uhrhammer, N. and Bignon, Y. J. (2008) 'Report of a family segregating mutations in both the APC and MSH2 genes: juvenile onset of colorectal cancer in a double heterozygote', *International Journal of Colorectal Disease*, 23(11), pp. 1131–1135. doi: 10.1007/s00384-008-0526-9.

Upadhyay, R. *et al.* (2013) 'Genetic polymorphisms in RNA binding proteins contribute to breast cancer survival.', *International journal of cancer. Journal international du cancer*, 132(3), pp. E128-38. doi: 10.1002/ijc.27789.

Vanharanta, S. *et al.* (2014) 'Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer Cancer Biology and Genetics Program , Memorial Sloan-Kettering Cancer MRC Cancer Unit , MRC / Hutchison Research Centre , University of Camb', *eLife*, pp. 1–24. doi: 10.7554/eLife.02734.

Varley, J. M. *et al.* (1999) 'Are there low-penetrance TP53 Alleles? evidence from childhood adrenocortical tumors.', *American journal of human genetics*, 65(4), pp. 995–1006. doi: 10.1086/302575.

Visscher, P. M. (2008) 'Sizing up human height variation', *Nature Genetics*, 40(5), pp. 489–490. doi: 10.1038/ng0508-489.

Visscher, P. M. *et al.* (2012) 'Five years of GWAS discovery', *American Journal of Human Genetics*. The American Society of Human Genetics, 90(1), pp. 7–24. doi: 10.1016/j.ajhg.2011.11.029.

Vogelaar, I. P. *et al.* (2017) 'Unraveling genetic predisposition to familial or early onset gastric cancer using germline whole-exome sequencing', *European Journal of Human Genetics*. Nature Publishing Group, (July), pp. 1–7. doi: 10.1038/ejhg.2017.138.

Volikos, E. *et al.* (2006) 'LKB1 exonic and whole gene deletions are a common cause of Peutz-Jeghers syndrome.', *Journal of medical genetics*, 43(5), pp. 4–6. doi: 10.1136/jmg.2005.039875.

Waldmann, J. *et al.* (2009) 'Screening of patients with multiple endocrine neoplasia type 1 (MEN-1): A critical analysis of its value', *World Journal of Surgery*, 33(6), pp. 1208–1218. doi: 10.1007/s00268-009-9983-8.

Wang, G. T. *et al.* (2014) 'Power analysis and sample size estimation for sequence-based association studies', *Bioinformatics*, p. btu296. doi: 10.1093/bioinformatics/btu296.

Wang, Q. *et al.* (2006) 'Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number', *Cancer*

Research, 66(12), pp. 6050–6062. doi: 10.1158/0008-5472.CAN-05-4618.

Warthin, A. S. (1913) 'Classics in oncology. Heredity with reference to carcinoma as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. By Aldred Scott Warthin. 1913.', *CA: a cancer journal for clinicians*, 35(6), pp. 348–59. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3931868.

Wasserman, J. D. *et al.* (2015) 'Prevalence and functional consequence of TP53 mutations in pediatric adrenocortical carcinoma: A children's oncology group study', *Journal of Clinical Oncology*, 33(6), pp. 602–609. doi: 10.1200/JCO.2013.52.6863.

Wei, Z. *et al.* (2012) 'Growth inhibition of human hepatocellular carcinoma cells by overexpression of G-protein-coupled receptor kinase 2.', *Journal of cellular physiology*, 227(6), pp. 2371–7. doi: 10.1002/jcp.22972.

Wellbrock, C. (2014) 'MAPK pathway inhibition in melanoma: resistance three ways: Figure 1', *Biochemical Society Transactions*, 42(4), pp. 727–732. doi: 10.1042/BST20140020.

Werling, U. and Schorle, H. (2002) 'Transcription factor gene AP-2γ essential for early murine development', *Molecular and cellular biology*, 22(9), pp. 3149–3156. doi: 10.1128/MCB.22.9.3149.

Whitworth, J., Skytte, A.-B., Sunde, L., Lim, D. H., *et al.* (2016) 'Multilocus Inherited Neoplasia Alleles Syndrome: A Case Series and Review.', *JAMA oncology*, 2(3), pp. 373–9. doi: 10.1001/jamaoncol.2015.4771.

Whitworth, J., Skytte, A.-B., Sunde, L., Lim, D. H., *et al.* (2016) 'Multilocus Inherited Neoplasia Alleles Syndrome', *JAMA Oncology*, 2(3), p. 373. doi: 10.1001/jamaoncol.2015.4771.

Wild, C. *et al.* (2015) 'Cancer risk: role of chance overstated.', *Science (New York, N.Y.)*, 347(6223), p. 728. doi: 10.1126/science.aaa6799.

Wu, S. *et al.* (2016) 'Substantial contribution of extrinsic risk factors to cancer development', *Nature*, 529(7584), pp. 43–47. doi: 10.1038/nature16166.

Xiao, W. *et al.* (2015) 'Conditional survival among patients with adrenal cortical carcinoma determined using a national population-based surveillance, epidemiology, and end results registry', *Oncotarget*, 6(42), pp. 44955–44962. doi: 10.18632/oncotarget.5831.

Yanaihara, N. *et al.* (2004) 'Reduced expression of MYO18B, a candidate tumor-suppressor gene on chromosome ARM 22Q, in ovarian cancer', *International Journal of Cancer*, 112(1), pp. 150–154. doi: 10.1002/ijc.20339.

Yao, F. *et al.* (2015) 'Recurrent Fusion Genes in Gastric Cancer: CLDN18-ARHGAP26 Induces Loss of Epithelial Integrity', *Cell Reports*, 12(2), pp. 272–285. doi: 10.1016/j.celrep.2015.06.020.

Zbuk, K. M. *et al.* (2007) 'Germline mutations in PTEN and SDHC in a woman with epithelial thyroid cancer and carotid paraganglioma', *Nature Clinical Practice Oncology*, 4(10), pp. 608–612. doi: 10.1038/ncponc0935.

Zhang, Y. *et al.* (2014) 'Role of Sphk1 in the malignant transformation of breast epithelial cells and breast cancer progression', *Indian Journal of Cancer*, 51(4). doi: 10.4103/0019-509X.175343.

Zheng, S. *et al.* (2016) 'Comprehensive pan-genomic characterization of adrenocortical carcinoma', *Cancer Cell*, 29(5), pp. 723–736. doi: 10.1016/j.ccell.2016.04.002.

Zhou, L. *et al.* (2016) 'G-protein-coupled receptor kinase 2 in pancreatic cancer: clinicopathologic and prognostic significance', *Human Pathology*. Elsevier Inc., 56, pp. 171–177. doi: 10.1016/j.humpath.2016.06.012.

11 Appendix

Germline pathogenic variants in *PALB2* and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without *CDH1* mutation: a whole-exome sequencing study

Eleanor Fewings, Alexey Larionov, James Redman, Mae A Goldgraben, James Scarth, Susan Richardson, Carole Brewer, Rosemarie Davidson, Ian Ellis, D Gareth Evans, Dorothy Halliday, Louise Izatt, Peter Marks, Vivienne McConnell, Louis Verbist, Rebecca Mayes, Graeme R Clark, James Hadfield, Suet-Feung Chin, Manuel R Teixeira, Olivier T Giger, Richard Hardwick, Massimiliano di Pietro, Maria O'Donovan, Paul Pharoah, Carlos Caldas, Rebecca C Fitzgerald, Marc Tischkowitz

Summary

Background **Germline pathogenic variants in the E-cadherin gene (*CDH1*) are strongly associated with the development of hereditary diffuse gastric cancer. There is a paucity of data to guide risk assessment and management of families with hereditary diffuse gastric cancer that do not carry a *CDH1* pathogenic variant, making it difficult to make informed decisions about surveillance and risk-reducing surgery. We aimed to identify new candidate genes associated with predisposition to hereditary diffuse gastric cancer in affected families without pathogenic *CDH1* variants.**

Methods **We did whole-exome sequencing on DNA extracted from the blood of 39 individuals (28 individuals diagnosed with hereditary diffuse gastric cancer and 11 unaffected first-degree relatives) in 22 families without pathogenic *CDH1* variants. Genes with loss-of-function variants were prioritised using gene-interaction analysis to identify clusters of genes that could be involved in predisposition to hereditary diffuse gastric cancer.**

Findings **Protein-affecting germline variants were identified in probands from six families with hereditary diffuse gastric cancer; variants were found in genes known to predispose to cancer and in lesser-studied DNA repair genes. A frameshift deletion in *PALB2* was found in one member of a family with a history of gastric and breast cancer. Two different *MSH2* variants were identified in two unrelated affected individuals, including one frameshift insertion and one previously described start-codon loss. One family had a unique combination of variants in the DNA repair genes *ATR* and *NBN*. Two variants in the DNA repair gene *RECQL5* were identified in two unrelated families: one missense variant and a splice-acceptor variant.**

Interpretation **The results of this study suggest a role for the known cancer predisposition gene *PALB2* in families with hereditary diffuse gastric cancer and no detected pathogenic *CDH1* variants. We also identified new candidate genes associated with disease risk in these families.**

Funding **UK Medical Research Council (Sackler programme), European Research Council under the European Union's Seventh Framework Programme (2007–13), National Institute for Health Research Cambridge Biomedical Research Centre, Experimental Cancer Medicine Centres, and Cancer Research UK.**

Lancet Gastroenterol Hepatol 2018

Published Online
April 26, 2018
http://dx.doi.org/10.1016/S2468-1253(18)30079-7

See Online/Comment
http://dx.doi.org/10.1016/S2468-1253(18)30120-1

Academic Laboratory of Medical Genetics (E Fewings MRes, A Larionov PhD, J Redman BSc, M A Goldgraben PhD, J Scarth BA, G R Clark PhD, M Tischkowitz FRCP), **Familial Gastric Cancer Study, Department of Oncology** (S Richardson RGN), **Centre for Cancer Genetic Epidemiology, Strangeway's Research Laboratory** (R Mayes HND, Prof P Pharoah FFPH), **Cancer Research UK Cambridge Institute** (J Hadfield PhD, S-F Chin PhD, Prof C Caldas FRCP), **and Medical Research Council (MRC) Cancer Unit, Hutchison/MRC Research Centre** (M di Pietro MD, Prof R C Fitzgerald FRCP), **University of Cambridge, Cambridge, UK; National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge, UK** (E Fewings, A Larionov, J Redman, M A Goldgraben, J Scarth, G R Clark, M Tischkowitz, Prof C Caldas, Prof R C Fitzgerald); **Department of Histopathology** (O T Giger PhD, M O'Donovan FRCPath) **and Department of Oesophago-Gastric Surgery** (R Hardwick FRCS), **Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; Precision Medicine and Genomics, Innovative Medicines and Early**

Introduction

Gastric cancer is the fourth most common cancer globally. The best characterised inherited gastric cancer is the diffuse type, which has the hallmark of multiple foci of signet-ring cells.[1] The term hereditary diffuse gastric cancer is used to describe families with a history of diffuse gastric cancer that meet the criteria of at least two cases of gastric cancer in first-degree or second-degree relatives regardless of age of onset (with one confirmed case of diffuse gastric cancer); one case of diffuse gastric cancer diagnosed before age 40 years; or a personal or family history of diffuse gastric cancer and lobular breast cancer, including one case diagnosed before age 50 years.[2,3]

Germline mutations in the E-cadherin gene (*CDH1*) explain 25–30% of hereditary diffuse gastric cancer cases, with more than 100 pathogenic germline variants currently described within this gene.[4] For families with hereditary diffuse gastric cancer and known pathogenic *CDH1* mutations, guidelines exist for risk assessment, disease management, surveillance (including regular endoscopies), and risk-reducing therapies (including prophylactic gastrectomy).[5,6] However, for families with no pathogenic variant in *CDH1*, the risk assessment is uncertain and, therefore, making decisions about and assessing the efficacy of risk-reducing strategies is challenging.

Development Biotech Unit, AstraZeneca, Cambridge, UK (J Hadfield); **Peninsula Clinical Genetics Service, Exeter, UK** (C Brewer MRCP); **West of Scotland Genetics Services, Glasgow, UK** (R Davidson MRCP); **Cheshire and Merseyside Regional Genetic Service, Liverpool, UK** (I Ellis FRCP); **Manchester Centre for Genomic Medicine, Manchester, UK** (Prof D G Evans FRCP); **Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, UK** (D Halliday FRCP); **Clinical Genetics Service, Guy's and St Thomas' NHS Foundation Trust, London, UK** (L Izatt FRCP); **West Midlands Regional Genetics Service, Birmingham, UK** (P Marks MSc); **Northern Ireland Regional Genetics Centre, Belfast City Hospital, Belfast, UK** (V McConnell MD); **Department of Gastroenterology, ZNA Jan Palfijn, Antwerp, Belgium** (L Verbist MD); **Department of Genetics, Portuguese Oncology Institute of Porto, Porto, Portugal** (Prof M R Teixeira MD); **and Institute of Biomedical Sciences, University of Porto, Porto, Portugal** (Prof M R Teixeira)

Correspondence to: Dr Marc Tischkowitz, Academic Laboratory of Medical Genetics, Addenbrooke's Treatment Centre, Cambridge Biomedical Campus, Cambridge CB2 2QQ, UK **mdt33@cam.ac.uk**

See Online for appendix

For more on the **1000 Genomes Project** see http://www.internationalgenome.org

For more on **Sorting Intolerant From Tolerant** see http://sift.bii.a-star.edu.sg

For more on **Polymorphism Phenotyping** see http://genetics.bwh.harvard.edu/pph2

Research in context

Evidence before this study

Knowledge of factors causing predisposition to hereditary diffuse gastric cancer in families with no pathogenic variants in *CDH1* is limited by the rarity of the disease, which makes doing large-scale association studies difficult. In 2015, Hansford and colleagues described variants in DNA repair-related genes in 144 families with hereditary diffuse gastric cancer without *CDH1* pathogenic variants. These genes included *PALB2*, *BRCA2*, and *ATM*, which are associated with breast cancer risk. Further investigation of these genes, other known cancer-predisposing genes, and genes associated with DNA repair will aid in the disease management of families with hereditary diffuse gastric cancer without *CDH1* pathogenic variants, whose risk of disease development is currently unknown.

Added value of this study

This study is one of the largest germline, whole-exome sequencing analysis of families with hereditary diffuse gastric cancer without *CDH1* mutation to date. Both affected and unaffected individuals were recruited from families with hereditary diffuse gastric cancer, providing the opportunity to look for protein-affecting variants that segregate with phenotype. We used a unique approach to pathway analysis that involved clustering of physically interacting genes that were enriched for variants in these families and annotating them with a Gene Ontology term. Additionally, combining findings from this study with data from previously published studies allowed a more complete analysis of the role of the cancer predisposition genes *PALB2* and *BRCA2*.

Implications of all the available evidence

We identified a cluster of interacting genes involved in DNA repair that could be associated with predisposition to hereditary diffuse gastric cancer, in particular, *PALB2*. These findings should help guide future studies seeking to elucidate the clinical implications of genes that have not been previously associated with hereditary diffuse gastric cancer. Identification of these genes could provide families with hereditary diffuse gastric cancer without *CDH1* pathogenic variants with improved information about the risks associated with their disease and allow them to make informed decisions about risk reduction and disease management.

Other familial cancer syndromes that have been linked to gastric cancer predisposition include Lynch syndrome, which is characterised by mutations in DNA mismatch repair genes; Peutz-Jeghers syndrome caused by mutations in *STK11*; and Li-Fraumeni syndrome, which is associated with germline *TP53* mutations.[2,7–9] Diffuse gastric cancer does not appear to be over-represented in these syndromes, although this association has not been comprehensively studied.

Predicted pathogenic variants in the DNA double-strand break repair genes *ATM*, *BRCA2*, and *PALB2* have been identified in several families with hereditary diffuse gastric cancer.[4,10] However, given the rarity of these variants, the associated risk of diffuse gastric cancer is hard to quantify, and these variants are not used in routine clinical testing to aid management of these families.

We aimed to identify new candidate genes for predisposition to hereditary diffuse gastric cancer in families without pathogenic *CDH1* variants.

Methods

Study design and participants

In this whole-exome sequencing study, we recruited 28 individuals diagnosed with diffuse gastric cancer and 11 unaffected relatives from 22 families with hereditary diffuse gastric cancer that had tested negative for *CDH1* pathogenic germline mutations as part of the Familial Gastric Cancer study (MREC 97/5/32) and for whom blood and tumour samples were available. Families (including first-degree and second-degree relatives) were categorised as having hereditary diffuse gastric cancer on the basis of existing criteria.[2,3,6]

Whole-exome sequencing and variant filtering

DNA was extracted from blood or saliva and prepared for 125-bp paired-end whole-exome sequencing using the Nextera Rapid Capture Exome Enrichment Kit (Illumina, San Diego, CA, USA). Sequencing was done on HiSeq-4000 or HiSeq-2500 platforms (Illumina, San Diego, CA, USA). Variant Call Format files were generated with a standard pipeline following Genome Analysis Toolkit (GATK) Best Practices recommendations for whole-exome data (appendix). The dataset was filtered to select uncommon (allele frequency <0·05 in the 1000 Genomes Project European sample) protein-affecting variants, including loss-of-function variants (stop site gained, stop site lost, start site lost, splice acceptor, splice donor, or frameshift), deleterious (predicted with Sorting Intolerant From Tolerant version 5.2.2) and damaging (predicted with Polymorphism Phenotyping version 2.2.2) missense variants, and inframe indels, that were observed in at least one of the 28 affected individuals. These filters were chosen to remove variants that were least likely to affect predisposition to hereditary diffuse gastric cancer. We considered the 11 unaffected family members separately on a per-family basis as a control group on which we did segregation analysis for identified candidate variants. We determined the allele frequency of all candidate variants in healthy controls (with no history of cancer) and allele counts in affected and unaffected individuals within families, and predicted downstream effects on the protein product.

Variants were aggregated into unique genes, which were then filtered to select those that contained at least one loss-of-function variant. We also removed the top 1%

most variable genes, which were identified by the number of rare, protein-affecting variants per gene. These genes typically possess many rare variants within the healthy population and, therefore, are unlikely to have a role in predisposition to hereditary diffuse gastric cancer or other diseases. Variant-filtering and gene-filtering steps are summarised in the appendix.

To analyse copy number variants, we applied the XHMM algorithm to the gene set, using principle component analysis to normalise read depth across exomes and a hidden Markov model to identify regions with variation in read depth.[11] Around a 50% decrease or increase in read depth was required for the variant to be considered for further analysis. Copy number variants were further explored in selected individuals with a CytoScan 750K genotyping array (Affymetrix, Santa Clara, CA, USA), according to the clinical protocol.

The results published here are in whole or part based on data generated by The Cancer Genome Atlas (TCGA), managed by the National Cancer Institute and the National Human Genome Research Institute. Controlled access data was requested and downloaded for the TCGA-STAD dataset, of which a subset of data from 88 cases of diffuse gastric cancer were analysed to further validate the role of the identified candidate genes in predisposition to diffuse gastric cancer.

Gene-interaction network analysis

Gene-interaction network analysis was used to identify variant-enriched candidate genes with interacting protein products; non-antagonistic, physically interacting proteins might have a similar effect on cell function and, therefore, might produce a shared phenotype when mutated. The filtered genes were put through the GeneMANIA Cytoscape plugin version 3.4.1, which places physically interacting genes into clusters.[12] A cluster was defined as a set of five or more physically interacting genes.

We used the PANTHER over-representation test (version 13.0) in the Gene Ontology Consortium enrichment analysis web-tool to assign Gene Ontology (GO) terms to clusters, applying the default Bonferroni correction for multiple testing.[13] Of the significant terms highlighted in the analysis, the most significant term that encompassed between ten and 200 genes was selected, consistent with previous studies.[14]

Allelic counts of all filtered, loss-of-function variants within the selected GO terms (regardless of GeneMANIA Cytoscape clustering) were aggregated and contingency tables were drawn. Variants were also aggregated for each GO term over a comparably filtered set of genes from 503 European individuals in phase 3 of the 1000 Genomes study (appendix).[15] We did a one-tailed Fisher's exact test using the R Stats package version 3.3.3 to test for enrichment of loss-of-function variants within each selected GO term in the families with hereditary diffuse gastric cancer compared with the 1000 Genomes European dataset. For this test, only one occurrence of a variant was counted per affected family. A link to the custom R scripts used for this analysis can be found in the appendix.

Validation by Sanger sequencing

Candidate variants were validated by Sanger sequencing. Germline DNA from blood and extracted tumour DNA were quantified with the Qubit dsDNA HS Kit (Invitrogen, Carlsbad, CA, USA), and custom flanking primers were designed for each variant (primer sequences are shown in the appendix). DNA fragments were amplified by PCR and the products were sequenced on an ABI Genetic Analyser (Applied Biosystems Foster City, CA, USA) with BigDye Terminator version 3.1 (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's instructions.

Tumour immunohistochemistry and microsatellite instability analysis

For more on **The Cancer Genome Atlas** see http://cancergenome.nih.gov

We used the Ventana MMR IHC Panel (Roche, Indianapolis, IN, USA) to do immunohistochemistry analysis of known mismatch repair genes in available tumours from individuals in which variants in mismatch repair genes were identified. The panel includes antibodies against MLH1, PMS2, MSH2, and MSH6.

To analyse microsatellite instability, 5-µm formalin-fixed, paraffin-embedded tumour sections were mounted on glass slides for dewaxing and manual microdissection. DNA was extracted with the QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany). We assessed the DNA for five standard microsatellite markers (BAT25, BAT26, NR21, NR24, and MONO27) using the Promega MSI Analysis System, version 1.2 (Promega, Madison, WI, USA). Poorly and moderately differentiated gastric tissue was compared with adjacent tumour-free tissue.

Analysis of *PALB2* and *BRCA2* variants in published studies

We searched PubMed without language restrictions between Jan 1, 2015, and Dec 31, 2017, using the term "hereditary diffuse gastric cancer" to identify sequencing studies reporting loss-of-function variants in *PALB2* and *BRCA2* in hereditary diffuse gastric cancer probands with no detected pathogenic *CDH1* variants. We included only publications released after the initial report[4] in 2015 of *PALB2* and *BRCA2* mutations associated with hereditary diffuse gastric cancer. For each of the four identified publications[4,10,16,17] and this study, we aggregated the allele counts of loss-of-function *PALB2* variants. The same counts were done across the 503 European samples from the 1000 Genomes Project and the 27 173 non-Finnish European individuals not in the TCGA from the Exome Aggregation Consortium (ExAC) control datasets.[18] We removed the well characterised, non-pathogenic *BRCA2* polymorphic stop codon in c.9976A→T from all datasets. We did a one-tailed Fisher's exact test using the R Stats package version 3.3.3 to test for enrichment of

	Number of samples sequenced*		Age of proband at diagnosis (years)	Cancers diagnosed in relatives of probands†	Candidate gene identified in proband
	Affected	Unaffected			
1	2	0	41	Diffuse gastric cancer (44)‡, gastric cancer (57)	None
2	2	4	27	Peritoneal cancer, ovarian cancer (22), diffuse gastric cancer (24)‡, diffuse gastric cancer (28)	None
3	1	0	40	Gastric cancer (28), diffuse gastric cancer (48)	None
4	1	2	55	Breast cancer, lung cancer, laryngeal cancer, gastric cancer, and diffuse gastric cancer (44, 52)	*PALB2*
5	1	0	36§	Diffuse gastric cancer (37), lung cancer (54), colorectal cancer (57), breast cancer (50), diffuse gastric cancer (61), diffuse gastric cancer (79), lung cancer (83)	None
6	1	0	37	Breast cancer, gastric cancer (63), gastric cancer (64)	*RECQL5*
7	2	0	36	Colorectal cancer, breast cancer (43), diffuse gastric cancer (55)‡	None
8	1	0	47¶	Diffuse gastric cancer (44)	*MSH2*‖
9	1	0	44	Diffuse gastric cancer (28)	None
10	1	2	28	Breast cancer, gastric cancer (44), gastric cancer (47)	None
11	4	1	28	Signet-ring cells‡, Signet-ring cells‡, breast cancer (40s), diffuse gastric cancer (45)‡, prostate cancer (60s), colorectal cancer (75)	*ATR*, *NBN*
12	1	0	68	Lung cancer, gastric cancer (49), gastric cancer (50), gastric cancer (76)	*MSH2*‖
13	1	0	47	Gastric cancer, gastric cancer (50s), gastric cancer (60s)	None
14	1	0	23	Diffuse gastric cancer (40s), diffuse gastric cancer (46), thyroid cancer (30)	None
15	1	0	53	Gastric cancer (49), gastric cancer (67), gastric cancer (71)	None
16	1	0	37	Gastric cancer, breast cancer (54), breast cancer (65), colorectal cancer (66)	None
17	1	0	45	Diffuse gastric cancer (42)	None
18	1	0	48	Gastric cancer (44), gastric cancer (54)	None
19	1	1	35	Lung cancer, uterine cancer (65)	None
20	1	0	55	Gastric cancer (51), colorectal cancer (76)	None
21	1	1	28	Gastric cancer (53), breast cancer (76), gastric cancer (80)	*RECQL5*
22	1	0	30	Gastric cancer, diffuse gastric cancer (67)	None

In total, 39 individuals were sequenced, including 11 unaffected relatives. Numbers in parentheses indicate age in years at diagnosis. *All probands were sequenced in this study. †Includes both first-degree and second-degree relatives; age in years at diagnosis is in parentheses when known. ‡Sequenced in this study. §This proband was also diganosed with colorectal cancer at age 47 years. ¶This proband was also diagnosed with lobular breast cancer at age 36 years. ||No microsatellite instability was detected in tumour.

Table 1: **Characteristics of 28 affected individuals in 22 families with hereditary diffuse gastric cancer without *CDH1* pathogenic variants**

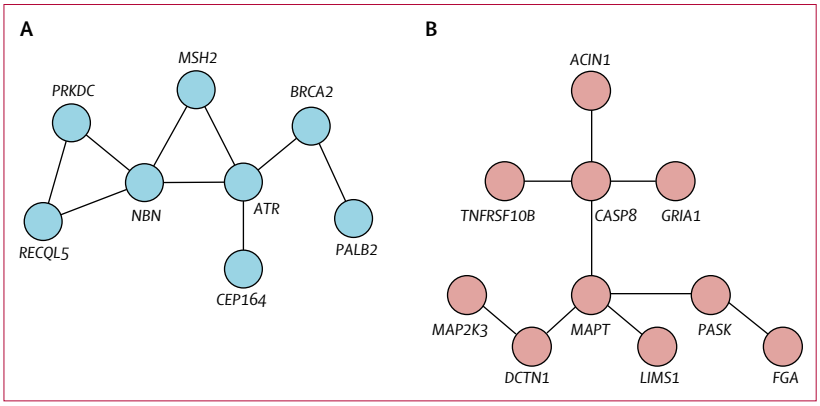


Figure 1: **Gene clusters identified via gene interaction analysis**
Lines indicate a physical interaction, as assigned by the GeneMANIA plugin for Cytoscape.[12] (A) Gene cluster to which the double-strand break repair GO term (GO:0006302) was assigned. (B) Gene cluster to which the negative regulation of extrinsic apoptotic signalling pathway via death domain receptors GO term (GO:1902042) was assigned. GO=Gene Ontology.

loss-of-function *PALB2* or *BRCA2* variants in the families with hereditary diffuse gastric cancer compared with either control dataset.

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the Article. EF, JR, and MT had access to the raw data. The corresponding author had full access to all of the data and the final responsibility for the decision to submit for publication.

Results

A whole-exome sequencing dataset of 39 individuals from 22 families with hereditary diffuse gastric cancer without pathogenic *CDH1* variants (table 1) was filtered to select 3973 uncommon, protein-affecting variants that were aggregated into 2847 genes. Exclusion of the top 1% of highly variable genes, and retention of genes with at least one loss-of-function variant in affected individuals, resulted in a set of 732 genes (1228 variants). Eight highly variable genes were excluded, including *ANKRD36*, *CDC27*, *HLA-DRB1*, *HLA-DRB5*, *MUC3A*, *MUC4*, *MUC6*, and *OR4C5*. Additionally, the presence of affected and unaffected family members within our dataset allowed for selection

	Number of sequenced individuals	Gene	Variant	Consequence	Protein change	Minimum allele frequency		SIFT	Polymorphism Phenotyping
						1000 Genomes European sample	ExAC non-TCGA European sample		
4	3	*PALB2*	c.757-758TAG→T	Frameshift deletion	Leu253fs	0	0	NA	NA
6	1	*RECQL5*	c.2806-2T→C	Splice-acceptor variant	NA	0	0	NA	NA
8	1	*MSH2*	c.967-968T→TCTCA	Frameshift insertion	Ser323fs	0	0	NA	NA
11	5	*ATR*	c.6075A→T	Stop site gain	Tyr2025X	0	0	NA	NA
11	5	*NBN*	c.1124+1G→C	Splice-donor variant	NA	0	0	NA	NA
12	1	*MSH2*	c.1A→C	Start site loss	Met1?*	0	0	Deleterious	Benign
21	2	*RECQL5*	c.2828C→T	Missense variant	Arg943His	0·002	0·014332	Deleterious	Probably damaging

SIFT=Sorting Intolerant From Tolerant. ExAC=Exome Aggregation Consortium. TCGA=The Cancer Genome Atlas. fs=frameshift. NA=not applicable. *Human Genome Variation Society nomenclature to indicate loss of a start site without experimental evidence of a new start site.[20]

Table 2: **Candidate variants in six families with hereditary diffuse gastric cancer without *CDH1* pathogenic variants**

of variants that segregated with phenotype on a per-family basis.

Gene-interaction network analysis of the 732 filtered genes identified two physical interaction clusters, to which GO terms were applied (figure 1). A cluster of eight genes was associated with the GO term double-strand break repair (GO:0006302; p<0·0001; figure 1A). A second cluster of ten genes was associated with the GO term negative regulation of extrinsic apoptotic signalling pathway via death domain receptors (GO:1902042; p=0·00517; figure 1B).

Loss-of-function variants within the filtered set of 1228 variants were aggregated under these two GO terms, including genes related to the GO terms that were not initially clustered by GeneMANIA. The double-strand break repair term was significantly enriched in families with hereditary diffuse gastric cancer compared with the 1000 Genomes European set (p=0·00051). By contrast, the apoptotic signalling pathway term was not enriched in families with hereditary diffuse gastric cancer (p=0·186), suggesting that the differences between the datasets in allele counts of DNA double-strand break repair genes cannot entirely be explained by technical differences that arise when using an externally produced control dataset.

Genes in the double-strand break repair GO term included *PALB2, MSH2, RECQL5, ATR*, and *NBN*, all of which were shown to be physically interacting in GeneMANIA (figure 1A). *BRCA2* was also a part of this set, but was disregarded from further study because it contained the well characterised, benign polymorphic stop codon c.9976A→T.[19]

Table 2 summarises the candidate variants. A heterozygous 2 bp frameshift deletion was identified in *PALB2* (c.757-758TAG→T [rs180177092, NM_024675.3]) in a patient from family 4 who was diagnosed with diffuse gastric cancer at age 55 years (figure 2). This loss-of-function variant at aminoacid position 253 is predicted to result in an early stop codon seven aminoacids downstream of the variant. Family 4 has a history of breast, lung, laryngeal, and diffuse gastric cancer (table 1, figure 2). Exome sequencing was also done on two unaffected siblings, one of whom also had the *PALB2* (c.757-758TAG→T [rs180177092, NM_024675.3]) variant. The affected proband had previously received treatment for *Helicobacter pylori* infection, but had tested negative at subsequent endoscopies.

Two heterozygous loss-of-function variants were identified in the mismatch repair gene *MSH2*: a start site loss (c.1A→C [rs267607911, NM_000251.2]) in a patient from family 12 and a frameshift insertion of 4 bp (c.967-968T→TCTCA [NM_000251.2]) in a patient from family 8 (table 2; appendix). Both families had a strong history of gastric cancer; however, only DNA from the probands was available for sequencing, so segregation analysis could not be done. Heterozygosity of both variants was maintained in tumour DNA from the patients, as confirmed by Sanger sequencing. Tumours from both probands showed normal expression of MSH2 and other mismatch repair proteins by immunohistochemistry compared with adjacent tumour-free tissue (appendix), and neither tumour showed evidence of microsatellite instability (appendix). Both probands with *MSH2* variants had previously tested negative for *H pylori*.

Heterozygous variants in the DNA repair genes *ATR* and *NBN*—a splice-donor variant (c.1124+1G→C [NM_002485.4]) in *NBN* and a predicted stop site-gain variant (c.6075A→T [NM_001184.3]) in *ATR*—were identified in the proband from family 11, who was diagnosed with diffuse gastric cancer at age 28 years (figure 3, table 2). Two siblings underwent risk-reducing gastrectomies, and subsequent pathological analysis of gastric tissue revealed the presence of signet-ring cells in both individuals. As such, these individuals were considered to be affected family members in this analysis. The father of the proband was diagnosed with diffuse gastric cancer at age 60 years and the mother had metastatic disease characterised by signet-ring cells,

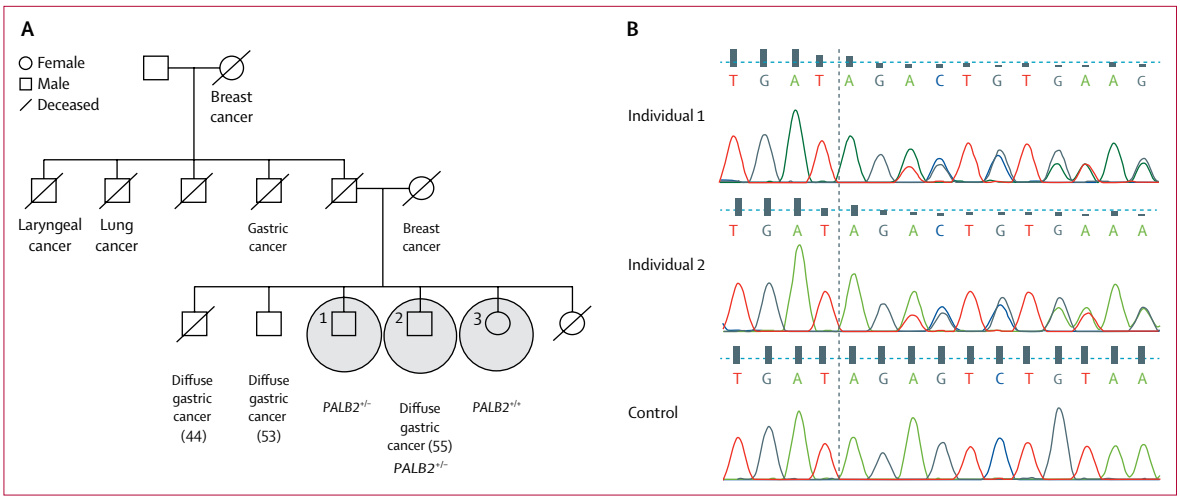


Figure 2: **Pedigree and cancer history for family 4**
(A) Whole-exome sequencing was done on the three circled individuals; age at diagnosis of cancer is shown in parentheses when known. (B) Chromatograms showing the *PALB2* frameshift variant (c.757-758TAG→T) in DNA from individuals 1 and 2 compared with control DNA.

which suggests that she had primary gastric cancer. All individuals in this family whose DNA was sequenced tested negative for *H pylori*. The proband and both siblings were heterozygous for loss-of-function variants in both *ATR* and *NBN*. The splice-donor variant in *NBN* was not seen in the father, and so was presumed to have been inherited maternally (DNA was only available for the father). The *ATR* variant was identified in the father, in whom no clinically relevant copy number variants were found. An unaffected, second-degree relative in family 11 did not have either variant.

Two variants were identified in the helicase gene *RECQL5* in different families. One was a missense variant (c.2828C→T [rs200535477, NM_004259.6]) in the proband from family 21, who was diagnosed with diffuse gastric cancer at age 28 years, and the other was a loss-of–function, splice-acceptor variant (c.2806-2T→C [rs201841487, NM_004259.6]) in the proband from family 6, who was diagnosed with diffuse gastric cancer at age 37 years (table 2; appendix). Both of these families included individuals across three generations who were diagnosed with gastric cancer and breast cancer. The proband of family 6 tested negative for *H pylori*, and the *H pylori* status of the proband from family 21 was unknown. DNA from the father of the proband in family 21 was sequenced, and the missense variant in *RECQL5* was not found.

We did not explore variants in other genes within the double-strand break repair cluster because the variants did not segregate with the disease in families containing affected and unaffected members.

We analysed data from previous studies[4,10,16,17] to estimate enrichment of loss-of-function variants in *PALB2* and *BRCA2* in families with hereditary diffuse gastric cancer. A loss-of-function variant (c.1438A→T [rs1057520653, NM_024675.3]) was identified and reported to us by collaborators (Teixeira MR, unpublished), however this variant was not included in the analysis because it did not fit our search criteria. Five (2%) of the 329 probands tested in these studies (including the present study) had loss-of-function *PALB2* variants (table 3). By contrast, *PALB2* variants were identified in 26 (<1%) of 27 173 individuals in the non-TCGA, non-Finnish European ExAC database ($p<0·0001$) and in one (<1%) of 503 individuals in the 1000 Genomes Project European database ($p=0·039$). Loss-of-function *BRCA2* variants were not enriched in families with hereditary diffuse gastric cancer compared with ExAC ($p=0·47$) or 1000 Genomes Project ($p=1·00$) individuals. No loss-of-function *PALB2* variants were identified in a set of 88 cases with sporadic diffuse gastric cancer from the TCGA.

Discussion

We found predicted pathogenic (protein-affecting) germline variants in known cancer-predisposing DNA repair genes (including *PALB2, MSH2, ATR, NBN*, and *RECQL5*) in six (27%) of 22 families with hereditary diffuse gastric cancer. This finding reflects the increasing number of cancer phenotypes found to be associated with existing cancer-predisposing genes as genomic analyses extend to rarer cancer subtypes. For example, mismatch repair genes were initially associated with increased risk of colorectal cancer, but were subsequently associated with risk of developing gastric and pancreatic cancers, among others.[21,22]

Simply identifying predicted pathogenic variants in known cancer-predisposing genes does not imply causality. For example, pathogenic germline variants in *MSH2* were not accompanied by altered expression of DNA mismatch repair proteins in tumour tissue in our study. To investigate causality, larger studies with

matched controls are required, which is not typically feasible for rare diseases. The generation of large control datasets, such as ExAC and 1000 Genomes, can be used to strengthen possible associations, as we have shown in the case of *PALB2*. When combining our data with those from previously published relevant studies,[4,10,16,17] including those in which no *PALB2* variants were found, we saw a significant over-representation of *PALB2* (but not *BRCA2*) pathogenic variants in families with hereditary diffuse gastric cancer compared with ExAC and 1000 Genomes controls; however, this finding was much less significant when comparing with the 503 Europeans in the 1000 Genomes Project set than when comparing with the 27 173 individuals from the non-Finnish, non-TCGA ExAC dataset.[15]

In several of the cases described by Sahasrabudhe and colleagues,[10] tumour molecular profiling was done and showed that carriers of *PALB2* mutations had mutational signatures indicative of defects in homologous recombination. PALB2 has an important role in homologous recombination during double-stranded DNA break repair through recruitment of BRCA2 and RAD51 to DNA breaks. Mutations in this gene are associated with an increased risk of breast and pancreatic cancers.[23–25] Even within families carrying *PALB2* mutations, cases of diffuse gastric cancer are likely to be rare and could be masked by a larger number of sporadic gastric adenocarcinomas, which means that associations with certain cancer subtypes might be missed in epidemiological studies of these families unless the pathology of all reported cancers is known. For example, a recent study[26] revealed that a rare serous subtype of endometrial cancer is over-represented in carriers of *BRCA1* variants, identifying a novel cancer association with a gene that has been intensively studied for more than 20 years.

ATR and *NBN* are also involved in initiating the response to double-strand DNA breaks. The *NBN* gene product (NBS1) associates with MRE11 and RAD50 to form a complex involved in the activation of the ataxia proteins ATM and ATR, which have roles in the recruitment of damage repair proteins, cell cycle regulation, and apoptosis. We identified single loss-of-function variants in *NBN* or *ATR* in the parents of the proband in family 11, both of whom had, or were suspected to have, diffuse gastric cancer. These variants were co-inherited in all three siblings in the family. Expression of either one of these variants might predispose family members to diffuse gastric cancer, although no other incidences of gastric cancer, and only one instance of late-onset prostate cancer, were noted in an extensive maternal and paternal family history. Slavin and colleagues[17] also identified a stop site-gain variant in *ATR* in an individual with intestinal-type adenocarcinoma and a strong family history of gastric cancer.

The unusual cancer pattern seen in family 11 (with all immediate family members [siblings and parents], but no extended family members, of the proband affected) might be attributed to multi-locus, inherited neoplasia alleles syndrome, in which inheritance of pathogenic mutations in multiple cancer-predisposing genes leads to an atypical or severe phenotype.[27] The close functional relationship between *NBN* and *ATR* in double-stranded

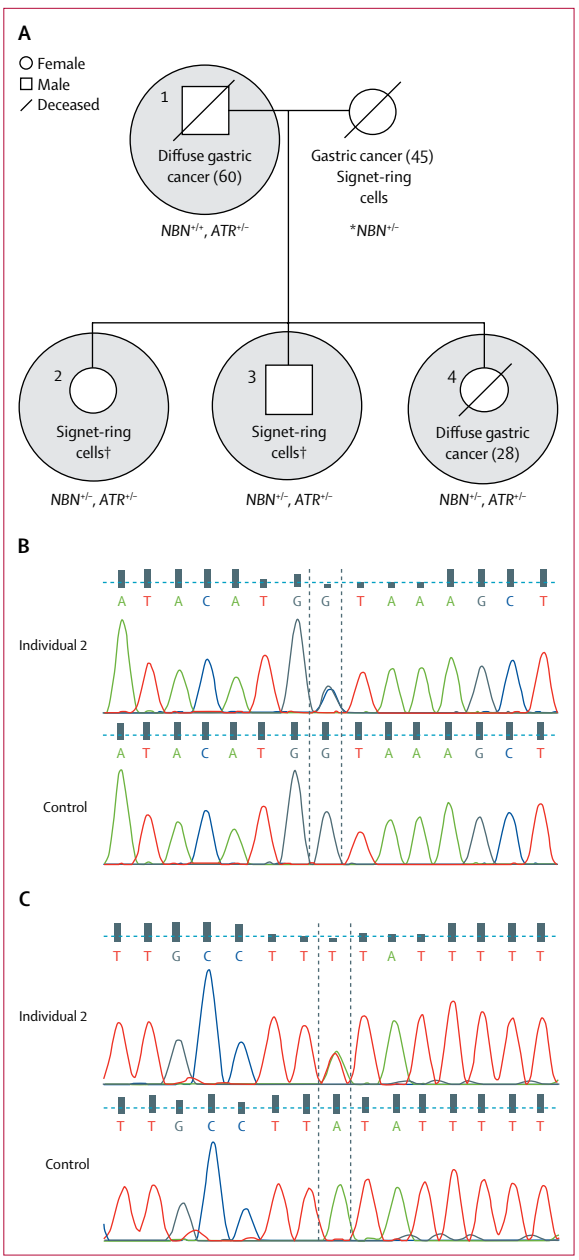


Figure 3: **Pedigree and cancer history of family 11**
(A) Whole-exome sequencing was done on the three circled individuals; the proband was patient 4. The presence of the variants in *NBN* (c.1124+1G→C) and *ATR* (c.6075A→T) among the four affected family members are shown. Age at diagnosis of cancer is shown in parentheses when known. *DNA was not available for the mother, and so the mother's genotype was assumed on the basis of the genotypes of the father and children. †These individuals underwent risk-reducing gastrectomies. (B) Chromatograms showing the *NBN* variants in DNA from individual 2 compared with control DNA. (C) Chromatograms showing the *ATR* variants in DNA from individual 2 compared with control DNA.

	Race	Patient ID	Diagnosis of proband (age at diagnosis in years)	Variant	Consequence	Protein change
Hansford et al (2015)[4]	European	P124	Diffuse gastric cancer (45)	c.1193AC→A	fs deletion	Val398fs
Sahasrabudhe et al (2017)[10]	European	CG-12	Intestinal gastric cancer (69)	c.1240C→T	Stop-site gain	Arg414Ter
Sahasrabudhe et al (2017)[10]	European	CG-008	Diffuse gastric cancer (48)	c.1240C→T	Stop-site gain	Arg414Ter
Sahasrabudhe et al (2017)[10]	European	GM037589	Gastric cancer (46)	c.1240C→T	Stop-site gain	Arg414Ter
Sahasrabudhe et al (2017)[10]	European	CG-05	Diffuse gastric cancer (50)	c.3201+1G→T	Splice-site variant	NA
Sahasrabudhe et al (2017)[10]	European	CG-039	Diffuse gastric cancer (47)	c.1882_1890delAAGTCCTGC	In-frame deletion	Lys628_Cys630del
Sahasrabudhe et al (2017)[10]	Latin American	CG-028	Intestinal gastric cancer (81)	c.1882_1890delAAGTCCTGC	In-frame deletion	Lys628_Cys630del
Sahasrabudhe et al (2017)[10]	Latin American	3CG-103	Mixed (79)	c.2753C→A	Missense	Pro918Gln
Fewings et al (this study)	European	GST_172_301	Diffuse gastric cancer (55)	c.757_758TAG→T	fs deletion	Leu253fs
Teixeira (unpublished)	European	GM048157	Diffuse gastric cancer (56)	c.1438A→T	Stop-site gain	Lys480Ter

None of the identified *PALB2* variants appeared in the 1000 Genomes Project European samples or in the Exome Aggregation Consortium European datasets. fs=frameshift. NA=not applicable.

Table 3: ***PALB2*** **variants identified in hereditary diffuse gastric cancer sequencing studies**

DNA break repair could indicate a potential combinatorial effect of variants in these genes, as potentially suggested by the young age of diagnoses or the presence of signet-ring cells in the siblings carrying these variants. By contrast, double heterozygosity of mutations in the DNA repair genes *BRCA1* and *BRCA2* in patients with breast cancer was found to be no more deleterious than a single heterozygous mutation.[28] Nevertheless, such double heterozygosity might have implications for genetic counselling that should be considered.

Genetic variants in genes involved in the mismatch DNA repair pathway are also associated with Lynch syndrome. A variant similar to the start-site loss variant that we identified in *MSH2* (c.1A→G) has previously been shown to have only a mild effect on protein function;[29] thus, this variant should not be treated as a typical loss-of-function variant. This attenuated effect on protein function could be due to the presence of an alternative start codon and a second non-mutated *MSH2* allele. Increased microsatellite instability, a measure of decreased MSH2 function, has been shown in patients and tumours with the c.1A→G start site-loss variant.[29] However, in tumours from both cases analysed here, MSH2 expression (detected by immunohistochemistry staining) was normal and no microsatellite instability was found. Although it is most likely that tumorigenesis was not caused by mismatch repair deficiency, we cannot rule out the possibility of a novel, non-mismatch repairmediated mechanism of carcinogenesis driven by variants in *MSH2*.

The helicase RECQL5 is important for prevention of aberrant homologous recombination and accumulation of double-strand DNA breaks, and thus for preservation of genome stability.[30] A missense *RECQL5* variant was identified in the proband from family 21, and was not found in the proband's unaffected father. A splice-acceptor variant in *RECQL5* was identified in an individual in family 6 who was diagnosed with diffuse gastric cancer at age 37 years. Both family 6 and family 21 had a history of breast and gastric cancer.

Previous studies have explored the role of known cancer predisposition genes in individuals with hereditary diffuse gastric cancer who do not have known *CDH1* mutations. Sahasrabudhe and colleagues[10] identified germline variants in *PALB2, BRCA1,* and *RAD51C* in families with diffuse gastric cancer. Hansford and colleagues[4] described variants in *ATM, BRCA2, MSR1,* and *STK11,* as well as a frameshift deletion in *PALB2*. This group has also uncovered a role for the *CDH1*-related adhesion gene *CTNNA1*. Although we did not find any variants of interest in *ATM, BRCA1, BRCA2, CTNNA1, MSR1, RAD51C,* or *STK11,* an exploration of *PALB2* variants in all families with hereditary diffuse gastric cancer sequenced in recent studies showed enrichment of loss-of-function variants in these families compared with control datasets. This finding makes a case for inclusion of *PALB2* in genetic testing for families with hereditary gastric cancer without *CDH1* mutations, and it is possible that individuals who carry *PALB2* mutations might benefit from platinum-based chemotherapy and treatment with PARP inhibitors.[31] However, the evidence is not yet sufficient to recommend surveillance of diffuse gastric cancer in carriers of *PALB2* mutations because the absolute risk is likely to be low in the absence of a family history.

Sporadic stomach cancers have been analysed as part of the TCGA study,[32] and an association was identified between truncating *PALB2* mutations and sporadic stomach adenocarcinoma. Of the individuals with sporadic stomach adenocarcinoma and *PALB2* mutations from the TCGA database, we selected 88 individuals with diffuse gastric cancer, as described by Bass and colleagues,[33] among which we did not identify any truncating *PALB2* variants. However, the average age at diagnosis for this cohort was 66 years, so this finding is perhaps not unexpected given the younger age of onset usually observed in hereditary cancers.

The rarity of patients with hereditary diffuse gastric cancer without pathogenic *CDH1* variants makes the collection of large datasets challenging. We used

gene-interaction network analysis to prioritise candidate gene variants that co-segregated with disease phenotype and were likely to be involved in predisposition to hereditary diffuse gastric cancer on the basis of knowledge of the biology of the disease. This approach did not overcome the problem of low statistical power due to a small sample size, which is often seen with rare cancer datasets, but it did allow for selection of the most plausible candidates from the available data.

We attempted an additional analysis of copy number variants within this dataset using the XHMM algorithm.[11] Although this analysis did not suggest any plausible candidates, at present, copy number variant analysis of germline whole-exome sequencing data is not validated and, therefore, some causal copy number variants could have been missed.

In summary, we found that rare, protein-affecting variants in DNA damage repair genes were enriched in families with hereditary diffuse gastric cancer without pathogenic *CDH1* variants compared with control datasets. Further studies of these genes in similar families are required to increase knowledge of the genetic basis of hereditary diffuse gastric cancer so that better informed decisions about risk reduction and management in affected family members can be made. Lastly, for many families with hereditary diffuse gastric cancer without pathogenic *CDH1* variants, the underlying cause remains unexplained even after whole-exome sequencing, and although whole-genome sequencing might identify some additional candidates in regulatory elements or structural variants, it seems unlikely that high-impact genes other than *CDH1* will be implicated in hereditary diffuse gastric cancer. Therefore, focusing on moderate-impact or low-impact cancer genes, such as *PALB2*, might be the way forward for future studies of genes associated with predisposition to disease in these patients.

Contributors

MT, PP, CC, and RCF conceived and designed the study. EF, AL, MAG, JS, JR, GRC, JH, and S-FC did the sequencing and data analyses. CB, RD, IE, DGE, DH, LI, PM, VMcC, LV, MRT, MdP, and RH recruited patients. SR and RM were responsible for coordinating recruitment, consent, and sample handling. MO'D and OTG provided histopathology input and reviewed reports and tumour samples. EF, AL, and PP did the statistical analyses and data interpretation. EF and MT wrote the manuscript and all authors reviewed the final version. MT is the guarantor.

Declaration of interests

DGE declares personal fees from AstraZeneca outside the submitted work. All other authors declare no competing interests.

Acknowledgments

This study was supported by UK Medical Research Council (Sackler programme), European Research Council (310018) under the European Union's Seventh Framework Programme (2007–13), National Institute for Health Research Cambridge Biomedical Research Centre, Experimental Cancer Medicine Centres, and Cancer Research UK. We thank the Human Research Tissue Bank, which is supported by the National Institute for Health Research Cambridge Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, UK. We also thank Tara Clancy (Manchester Centre for Genomic Medicine, Manchester, UK), Cecilia Compton (Clinical Genetics Service, Guy's and St Thomas' NHS Foundation Trust, London, UK), Sarah Everest (Peninsula Clinical Genetics Service, Exeter, UK), Vicky Hunt (Peninsula Clinical Genetics Service), Emma Kivuva (Peninsula Clinical Genetics Service), Anna Lehmann (Clinical Genetics, St George's University Hospitals NHS Foundation Trust, London, UK), Mark Longmuir (West of Scotland Genetics Services, Glasgow, UK), Ana Peixoto (Department of Genetics, Portuguese Oncology Institute of Porto, Porto, Portugal), Peter Risby (Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, UK), and Sarah Rose (Clinical Genetics Service, Guy's and St Thomas' NHS Foundation Trust) for their help in recruiting families and the mutation analysis.

References

1 Guilford P, Blair V, More H, Humar B. A short guide to hereditary diffuse gastric cancer. *Hered Cancer Clin Pract* 2007; **5:** 183–94.
2 van der Post RS, Vogelaar IP, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline *CDH1* mutation carriers. *J Med Genet* 2015; **52:** 361–74.
3 Caldas C, Carneiro F, Lynch HT, et al. Familial gastric cancer: overview and guidelines for management. *J Med Genet* 1999; **36:** 873–80.
4 Hansford S, Kaurah P, Li-Chang H, et al. Hereditary diffuse gastric cancer syndrome: *CDH1* mutations and beyond. *JAMA Oncol* 2015; **1:** 23–32.
5 Mi EZ, Mi EZ, di Pietro M, et al. Comparative study of endoscopic surveillance in hereditary diffuse gastric cancer according to *CDH1* mutation status. *Gastrointest Endosc* 2018; **87:** 408–18.
6 Fitzgerald RC, Hardwick R, Huntsman D, et al. Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *J Med Genet* 2010; **47:** 436–44.
7 Sereno M, Aguayo C, Guillén Ponce C, et al. Gastric tumours in hereditary cancer syndromes: clinical features, molecular biology and strategies for prevention. *Clin Transl Oncol* 2011; **13:** 599–610.
8 van Lier MGF, Westerman AM, Wagner A, et al. High cancer risk and increased mortality in patients with Peutz-Jeghers syndrome. *Gut* 2011; **60:** 141–47.
9 Masciari S, Dewanwala A, Stoffel EM, et al. Gastric cancer in individuals with Li-Fraumeni syndrome. *Genet Med* 2011; **13:** 651–57.
10 Sahasrabudhe R, Lott P, Bohorquez M, et al. Germline mutations in *PALB2, BRCA1,* and *RAD51C*, which regulate DNA recombination repair, in patients with gastric cancer. *Gastroenterology* 2017; **152:** 983–86.
11 Fromer M, Purcell SM. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr Protoc Hum Genet* 2014; **81:** 7.23. 1–21.
12 Montojo J, Zuberi K, Rodriguez H, et al. GeneMANIA cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* 2010; **26:** 2927–28.
13 Blake JA, Christie KR, Dolan ME, et al. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015; **43:** D1049–56.
14 Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* 2017; **49:** 1767–78.
15 Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature* 2015; **526:** 68–74.
16 Vogelaar IP, van der Post RS, van Krieken JHJ, et al. Unraveling genetic predisposition to familial or early onset gastric cancer using germline whole-exome sequencing. *Eur J Hum Genet* 2017; **25:** 1246–52.
17 Slavin T, Neuhausen SL, Rybak C, et al. Genetic gastric cancer susceptibility in the International Clinical Cancer Genomics Community Research Network. *Cancer Genet* 2017; **216–17:** 111–19.
18 Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60706 humans. *Nature* 2016; **536:** 285–91.
19 Higgs JE, Harkness EF, Bowers NL, et al. The *BRCA2* polymorphic stop codon: stuff or nonsense? *J Med Genet* 2015; **52:** 642–45.
20 Human Genome Variation Society. Discussions regarding the description of sequence variants. http://www.hgvs.org/mutnomen/disc.html#Met (accessed April 16, 2018).
21 Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. Familial gastric cancer: genetic susceptibility, pathology, and implications for management. *Lancet Oncol* 2015; **16:** e60–70.

22 Lynch HT, Voorhees GJ, Lanspa SJ, McGreevy PS, Lynch JF. Pancreatic carcinoma and hereditary nonpolyposis colorectal cancer: a family study. *Br J Cancer* 1985; **52:** 271–73.

23 Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in *PALB2*. *N Engl J Med* 2014; **371:** 497–506.

24 Pauty J, Rodrigue A, Couturier A, Buisson R, Masson J-Y. Exploring the roles of PALB2 at the crossroads of DNA repair and cancer. *Biochem J* 2014; **460:** 331–42.

25 Easton DF, Pharoah PDP, Antoniou AC, et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 2015; **372:** 1–15.

26 Saule C, Mouret-Fourme E, Briaux A, et al. Risk of serous endometrial carcinoma in women with pathogenic *BRCA1/2* variant after risk-reducing salpingo-oophorectomy. *J Natl Cancer Inst* 2018; **110:** 2017–19.

27 Whitworth J, Skytte A-B, Sunde L, et al. Multilocus inherited neoplasia alleles syndrome. *JAMA Oncol* 2016; **2:** 373.

28 Leegte B. Phenotypic expression of double heterozygosity for *BRCA1* and *BRCA2* germline mutations. *J Med Genet* 2005; **42:** e20.

29 Kets CM, Hoogerbrugge N, van Krieken JHJM, Goossens M, Brunner HG, Ligtenberg MJL. Compound heterozygosity for two *MSH2* mutations suggests mild consequences of the initiation codon variant c.1A>G of *MSH2*. *Eur J Hum Genet* 2009; **17:** 159–64.

30 Saponaro M, Kantidakis T, Mitter R, et al. RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell* 2014; **157:** 1037–49.

31 Ledermann JA, Kohn EC. PARP inhibitors for *BRCA1/2* mutation-associated and *BRCA*-like malignancies. *Ann Oncol* 2014; **25:** 32–40.

32 Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun* 2015; **6:** 10086.

33 Bass AJ, Thorsson V, Shmulevich I, et al. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014; **513:** 202–09.

PALB2 as a familial gastric cancer gene: is the wait over?

KTSDesign/Science Photo Library

Lancet Gastroenterol Hepatol 2018

Published Online
April 26, 2018
http://dx.doi.org/10.1016/S2468-1253(18)30120-1

See Online/Articles
http://dx.doi.org/10.1016/S2468-1253(18)30079-7

Gastric cancer is the third leading cause of cancer-related mortality, accounting for more than 730 000 deaths worldwide each year.[1] The prognosis of patients with gastric cancer is dismal because most gastric tumours are diagnosed in late stages when 5-year survival is less than 20%.[1] Whereas gastric tumours with intestinal histologies are linked to a history of *Helicobacter pylori*-associated gastritis, the natural history of diffuse gastric cancers is less well understood.[1] New tools for the prevention and early detection of early gastric cancer are needed to improve survival outcomes. Around one in ten patients with gastric cancer report a family history of malignancy, suggesting an underlying genetic predisposition.[1] Thus, genetic information could be used for gastric cancer prevention.

Until recently, *CDH1* was the only known gene associated with familial gastric cancer.[2] *CDH1* mutations frequently cause hereditary diffuse gastric cancer syndrome, and detection of mutations in this gene is routinely used for risk assessment and disease management.[1,2] However, more than 40% of families with hereditary diffuse gastric cancer syndrome do not carry *CDH1* mutations, suggesting the existence of additional predisposing genes.[3,4] In the past 3 years, our group and others have suggested that mutations in homologous recombination DNA repair genes, such as *PALB2*, probably explain a substantial proportion of inherited gastric cancers.[3,5] The study by Eleanor Fewings and colleagues[6] in *The Lancet Gastroenterology & Hepatology* further supports a role for *PALB2* in predisposition to hereditary gastric cancer.

To identify new gastric cancer-associated genes, the investigators performed whole-exome sequencing in 22 families with hereditary diffuse gastric cancer syndrome that had previously tested negative for pathogenic variants in *CDH1*. Exome data from affected family members was used to prioritise candidate genes with predicted loss-of-function variants using interaction analyses. A cluster of DNA repair genes, which included known cancer-associated genes such as *MSH2* and *PALB2*, was identified as being significantly enriched for loss-of-function variants. Subsequent analyses of this cluster provided evidence of co-segregation of loss-of-function variants with disease in six (27%) of the 22 families. A frameshift *PALB2* variant was identified in one family, and loss-of-function variants in both *ATR* and *NBN* were identified in another family. *RECQL5* mutations were identified in two families; two others had variants in *MSH2*, a gene known to be associated with Lynch syndrome.[7] Tumour analyses in the individuals with *MSH2* variants did not reveal microsatellite instability or loss of mismatch repair protein expression (both hallmarks of Lynch syndrome[7]), suggesting that these variants were either non-pathogenic or that a new mechanism leads to gastric cancer in these families. To exclude a possible environmental cause, the family members were assessed for history of *H pylori* infection, the strongest known risk factor for gastric cancer,[1] and most mutation carriers had previously tested negative for the bacteria. The investigators also combined their data with those from two previous reports[3,5] and found that *PALB2* loss-of-function variants were substantially more common in families with hereditary gastric cancer than in the general population.

Genetic studies in gastric cancer have lagged behind those in other gastrointestinal cancers, such as colorectal cancer, for which several susceptibility genes have been reported.[8] The study of gastric cancer genetics is difficult in part because the high mortality makes large family studies unfeasible. Furthermore, dissimilar to colorectal cancer, gastric cancer is more commonly diagnosed in low-income countries, where research and clinical practice of cancer genetics is limited by funding and the scarcity of medical professionals in the field, among other reasons.[9]

Therefore, the investigators should be commended for these findings. Nevertheless, given the small number of families with *PALB2* variants (only ten have been reported thus far), translating these findings into prevention of gastric cancer will require additional research. For example, should cases of familial intestinal gastric cancer be tested for *PALB2* variants? The study[5] from our group identified seven individuals with *PALB2* mutations, of whom three had intestinal tumours and one had a tumour of mixed histology. Given that hereditary diffuse gastric cancer syndrome includes only cases with diffuse histology, our findings suggest that *PALB2* mutations define a unique syndrome with heterogeneous histological types.

Further studies focusing on non-diffuse gastric cancer are therefore needed to understand whether *PALB2* mutations define a familial gastric cancer syndrome.

Germline mutations in *PALB2* have already been associated with an increased risk of breast and pancreatic cancer, and hundreds of families are known to have mutations in this gene.[10] It is now important to identify which factors increase the risk of gastric cancer in such families. Both the study by Fewings and colleagues[6] and our study[5] found insufficient evidence of a modifier role for *H pylori* infection in *PALB2* mutation carriers, but it is possible that the risk of gastric cancer is affected by *PALB2* mutation hotpots, modifier genes, and other unaccounted environmental factors. These questions highlight the need for further research.

This emerging body of data provides a compelling case for offering *PALB2* testing to families with hereditary gastric cancer that have previously tested negative for pathogenic *CDH1* mutations. Given that the tumours of individuals with *PALB2* mutations are likely to be deficient in homologous recombination DNA repair (as shown in our study[5]), these individuals could potentially benefit from PARP inhibitor therapies, which might be effective in homologous DNA repair-deficient tumours.[11] 20 years after the discovery of *CDH1*, the identification of *PALB2* as a new gene potentially associated with familial gastric cancer is of dual importance for both prevention and treatment of gastric cancer.

Luis G Carvajal-Carmona
Genome Center and Department of Biochemistry and Molecular Medicine, University of California, Davis, CA 95616, USA
lgcarvajal@ucdavis.edu

I declare no competing interests. I receive research funding from the National Cancer Institute (R01CA223978 and R21CA199631) of the National Institutes of Health. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

1 Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. *Cancer Epidemiol Biomarkers Prev* 2014; **23:** 700–13.
2 Guilford P, Hopkins J, Harraway J, et al. E-cadherin germline mutations in familial gastric cancer. *Nature* 1998; **392:** 402–05.
3 Hansford S, Kaurah P, Li-Chang H, et al. Hereditary diffuse gastric cancer syndrome: *CDH1* mutations and beyond. *JAMA Oncol* 2015; **1:** 23–32.
4 van der Post RS, Vogelaar IP, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline *CDH1* mutation carriers. *J Med Genet* 2015; **52:** 361–74.
5 Sahasrabudhe R, Lott P, Bohorquez M, et al. Germline mutations in *PALB2*, *BRCA1*, and *RAD51C*, which regulate DNA recombination repair, in patients with gastric cancer. *Gastroenterology* 2017; **152:** 983–86.
6 Fewings E, Larionov A, Redman J, et al. Germline pathogenic variants in *PALB2* and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without *CDH1* mutation: a whole-exome sequencing study. *Lancet Gastroenterol Hepatol* 2018; published online April 26. http://dx.doi.org/10.1016/S2468-1253(18)30079-7.
7 Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. *Nat Rev Cancer* 2015; **15:** 181–94.
8 Carvajal Carmona LG, Tomlinson I. The hunting of the snark: whither genome-wide association studies for colorectal cancer? *Gastroenterology* 2016; **150:** 1528–30.
9 Rastogi T, Hildesheim A, Sinha R. Opportunities for cancer epidemiology in developing countries. *Nat Rev Cancer* 2004; **4:** 909–17.
10 Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in *PALB2*. *N Engl J Med* 2014; **371:** 497–506.
11 Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat Commun* 2015; **6:** 8683.